CONSCIOUSNESS IN SCIENTIFIC AND FOLK PSYCHOLOGY.

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Master of Science in Psychology

in the

University of Canterbury

by

Anthony P. Atkinson

University of Canterbury

1993

Contents.

Chapter 3.
Functionalism in action: cognitive models of consciousness.

Chapter 4.
Problems and prospects: Qualia and epiphenomenalism.

Acknowledgements.

I am indebted to several friends for their conversation, advice, and instruction over the past two years. I am especially grateful to Gill Rhodes, Paul Russell, and Simon Kemp for tutoring me in cognitive psychology, Jack Copeland for introducing me to the philosophy of cognitive science, Brian Haig for showing me the value and scope of philosophical psychology, and Derek Browne for introducing me to the philosophy of mind, and for his philosophical and editorial advice.

Most of all I would like to thank my supervisor, Ken Strongman, for his unflinching support and encouragement.

Abstract.

The intentional properties and subjective qualities of conscious states pose special problems for physicalism. Yet 'consciousness' is a term of the vernacular that picks out such a heterogeneous group of phenomena that it will not be a good explanandum for science. This thesis adopted the position that we are licensed to theorize about the phenomena of consciousness, provided we are careful to dump all excess folk-psychological baggage surrounding the term. It was argued that the purposes and goals of folk psychology differ considerably from those of scientific psychology, for folk psychology is first and foremost a *craft*.

Cognitive psychology is bound to the analytical strategy by way of *functionalism*. Various forms of functionalism were investigated, and two non-mutually exclusive versions were favoured: *homuncular functionalism* and *microfunctionalism*. This led to the view that nature is multi-levelled, and therefore that functionalism may be better known as *structural-functional theory*. *S-F* theory should seek to explain the processes and structures of the mind-brain, rather than attempt to find the states posited by folk psychology within the cognitive system.

Traditional cognitive models view the mind as a highly structured system of semi-autonomous processors under the monitoring and guidance of a central executive. But this thesis argued that to postulate a 'consciousness module', while a natural extension of functionalist 'boxology', is merely to pander to our folk-psychological intuitions of the will or 'inner self'. Some of the 'new wave' of cognitive models — those that do not posit an executive — were reviewed.

Phenomenal consciousness is the one major stumbling block for

physicalist theories. Although this thesis agreed that qualia do not exist, it was evident that no theory has yet provided a bridge across the explanatory gap between third-person science and first-person phenomenology over which sceptics feel safe to cross. Nevertheless, it was argued that Dennett's (1991a) latest theory, with its intelligent use of metaphors and analogies, is one of the most promising steps in the right direction.

Finally, it was argued throughout that an interdisciplinary approach is crucial if science is to uncover the mysteries of consciousness.

CHAPTER 1.

PHYSICALISM, FOLK vs. SCIENTIFIC PSYCHOLOGY, AND CONSCIOUSNESS.

Part I. Introduction.

1. *Science today: physicalism.*

Physicalism is the general doctrine underlying science today, and may best be viewed as relative to the physics of the day (Armstrong, 1987); that is, as providing the fundamental assumptions upon which our theories of nature are based. Physicalism admits of a wide range of different stances on the nature of the world, and particularly on the mind-body relationship. Generally physicalism is taken to be the position that mental states, processes, and the like *are the same as* (just are) physical states and processes.

Thus the primary concern of physicalism is the use of the language and theories of the physical sciences in the description, explanation, and prediction of the purposeful behaviour of humans and animals, and as such poses the question of whether "we need to make ineliminable reference to [mental phenomena] when explaining behaviour." (Wilkes, 1978, p.10). Or more explicitly, physicalism for the brain and behavioural sciences can be described as: "the attempt to correlate explanations of actions couched in psychological terms with descriptions and explanations of cerebral states, events, and processes couched in neurophysiological terms." (Wilkes, 1978, p.29). There are two points to note about this statement: (i) as Wilkes readily admits, the terms 'correlate' and 'explain' are still quite vague, but that is not such a bad thing, for (ii), as Wilkes goes on to suggest, this vagueness can be put to advantage,

because it does not commit us to reductionism.

Generally speaking, then, physicalism is the conception of the world as a physically closed system, i.e., every event is explicable in purely physical terms; there is nothing but the physical world. Just what this claim amounts to for the nature and progress of science is a matter of considerable debate. In the eyes of a stereotypical reductionist, for instance, physicalism involves two types of theories: psychology and neurophysiology. On this view, neurophysiology is regarded as the more basic or fundamental science "because it explains *why* the laws of psychology hold to the extent they do, something that psychology itself cannot explain; and because it is believed to be reducible to the science we consider the most fundamental: physics." (Wilkes, 1978, p.30). Reductionism is the perfect bedfellow for monistic physicalism. A reductionist stance is not the only position open to the physicalist, however. Indeed, if nature is viewed as multi-levelled (as I discuss in chapter 2), then our theories of nature are also likely to be multi-levelled (chapter 5).

2. *Physicalism and explanations of the mental.*

All conscious mental states are generally considered to exhibit one or other, or both, of two characteristics (e.g., Rosenthal, 1986): *intrinsic properties* — what the sensations, emotions, etc., *feel like* to the person who is experiencing them; and the *meaning* or *content* of those experiences — what they are about or represent. The conscious mental states picked out by their phenomenal characteristics are typically "the sensation-type mental phenomena: pains, itches, tingles, twitches, images, after-images, sense impressions." (Wilkes, 1978, p.11). The conscious mental states picked out by

their representational content are typically the propositional attitudes: beliefs, desires, memories, thoughts, and the like.

The physicalist strategy is to seek explanations of the mental in terms of the physics and chemistry of the brain; to find the neurobiological substrate of consciousness and other mental phenomena. If the above characterization of conscious mental states is more or less correct, then a major challenge for the physicalist is to explain, in purely physical terms, the meaningful content of our thoughts, beliefs, and desires, and the apparent intrinsic qualities of our sensory experience. But how is this even conceivable? Physicalism seems to run counter to our everyday, common-sense view of our mental life, for the latter "is pretty much Cartesian-dualist through and through" (Dennett, 1991c, p.137). Science thus confronts a paradox: from our own perspective at least, our conscious mental states appear to be so far removed or so utterly different from anything else in the physical world (from 'movements of atoms in the void') that we are pushed to the intuition that indeed they are like nothing else in the physical world, and even if they are part of that physical world, then they will certainly not yield to any physical explanation (see discussion of McGinn, 1989, 1991, in section 2.2 of part II, below). From the perspective of the person on the street, mental life is the last bastion of human uniqueness, safe in its inner sanctum from the ever advancing physical sciences. This is how Dennett (1991a) illustrates the problem:

> "How could any combination of electrochemical happenings in my brain somehow add up to the delightful way those hundreds of twigs geneflucted in time with the music? How could some information-processing event in my brain be the delicate warmth of the sunlight I felt falling on me? For that matter, how could an event in my brain be my sketchily visualized mental image of ... some other information-processing event in my brain? It does seem impossible." (pp.26-27).

The concept of consciousness is one aspect of mind that poses special problems for physicalism: it appears to elude or defy adequate scientific explanation, remaining even more mystifying than probably any other mentalistic concept. Yet most of us would regard conscious experience as the essence of what it is to be human. Most would agree, for example, that in our usual waking state we are in some way subjectively aware of our physical and social environment and our interaction with it, that our sensory experience of the physical world is imbued with all manner of intricate and colourful qualities, that we have a sense of self, and that we have a variety of thoughts (seemingly linguistic, visual, or auditory), beliefs, desires, intentions, emotions, etc. Furthermore, it is commonly held that these mental states have special characteristics or qualities, and that these qualities can be described, although often with some difficulty (what Dennett, 1988, calls the apparently "ineffable, intrinsic, private, directly or immediately apprehensible" quality of our experience). Nagel (1974/ 1979, 1986), characterizes this special quality of mental experience as the 'what it is like to be X' (see chapter 4). A frequently and sometimes vehemently held corollary of this account of the special nature of human consciousness is that it appears that it cannot be captured in the language of the physical sciences.

How then is the physicalist to reconcile the everyday intuitions about our mental life with her commitment to explaining the entire physical world, of which conscious mental states are presumed to be a part, in solely physical terms? When the physicalist steps back from the task at hand she is faced with two overriding general questions: What is it about these apparently essential characteristics of human nature that makes it so hard for science to get a handle on them?; and: How do we best view the mental realm, and hence how do we

best approach the scientific study and explanation of consciousness? These are the motivating questions for the present thesis. By concentrating my efforts on the latter question, I also hope to shed some light on the former.

A key problem for any scientific description and explanation of consciousness is how it is possible to reconcile the subjective nature of experience with physicalist theories of mind. Some argue (e.g., Nagel, 1974, 1979, 1986) that the subjective character of conscious experience will forever lie outside the boundaries of objective science. However (as I will attempt to show in chapter 4) the arguments advanced in support of such claims are either faulty or not as significant as they initially appear. Nevertheless, the subjective nature of experience, the "phenomenal mind" (Jackendoff, 1987), still presents the biggest stumbling-block for functionalist and physicalist theories of mind-brain and behaviour. The other major stumbling-block concerns the status of intensional language in scientific — physicalist or functionalist — descriptions and explanations of human action and behaviour. This issue, which is bound up with the notion of 'intentionality' as a mark of the mental, is a topic of considerable unresolved debate amongst philosophers of mind (see Part II of the present chapter).

Despite these objections, there is a growing body of assorted scientists and philosophers who are making significant inroads into the scientific explanation and understanding of consciousness and conscious phenomena. While neuroscience is undoubtedly a crucial part of this scientific push into explaining all we mean when we speak of consciousness (see e.g., P. S. Churchland, 1983, 1988), there are many (including myself) who believe that it is not the only part. This thesis will examine some other major contenders for that elusive prize: our best explanation of the mental.

There are a number of opinions, arguments, and theses as to why consciousness should and needs to come under the legitimate realm of scientific inquiry. The predominant thesis underpinning these views, at least from the point of view of psychology, is a functionalist view of the mind-brain (chapter 2). It is the cognitive sciences that have taken the functionalist perspective to heart, and developed it into detailed theses of mind (chapter 3). Despite the promise of some computational cognitive theories, however, it is not apparent that they deal adequately with the problems posed by the subjective nature of conscious experience (chapter 4). But what may amount to a deeper concern with such functionalist theories of consciousness is that, when taken to their logical end points, they tend to postulate entities and functions that appeal to a ubiquitous, yet outmoded and seriously mistaken view of the mind: "Cartesian materialism" (Dennett, 1991a; Dennett & Kinsbourne, 1992; see chapters 3, 4, 5). Finally it will be seen that functionalist theories, while remaining "respectable, useful and probably necessary", to borrow a phrase from Mandler (1975), cannot be the one and only vehicles for our best explanations of all that falls under the rubric 'consciousness' (chapter 5).

Part II: Folk craft and scientific explanation.

1. *Folk psychology, scientific psychology, and consciousness.*

1.1 *The status of folk psychology.*

Folk psychology is literally what it says it is: the everyday psychology of folk; the common-sense practice of describing, explaining and predicting our own and other persons' everyday behaviours and mental states in terms of beliefs, desires, thoughts, intentions, emotions, and the like. In a nutshell, folk psychology encompasses all commonsense concepts of our mental life; and 'consciousness' is a concept of folk psychology *par excellence.*

One central question in the philosophy of mind — indeed, it is also a central question for any science of the mind — concerns how this everyday competence squares with our general physicalist view of the world. This debate is concerned with the concept of intentionality and the status of folk psychology — what place, if any, does it have in a mature science; a debate which is too involved and lengthy for extended treatment in the present paper. (The literature here is extensive; for good overviews of the issues, see e.g., P. M. Churchland, 1981; Clark, 1989; Dennett, 1987; Fodor, 1985; Lyons, 1990, 1991; Sterelny, 1990; Stich, 1983; Wilkes, 1978.) Nevertheless, my views on the debate will emerge in what follows, for a discussion of the nature of folk psychology and its relationship with scientific psychology is central to my thesis.

One major group of players in this debate are the eliminativists, who hold that the advance of the physical sciences will eventually eliminate all folk-psychological talk. Paul M. Churchland (e.g., 1981, 1988), for example, takes neuroscience to be the supplanting doctrine, whereas Stich (1983) takes it to be some form of computational psychology. According to these arguments, folk psychology — that is, any scientific practice openly embracing folk-

psychological talk — is a degenerative research program. This rests partly on the claim that folk psychology is 'theoretical', in some systematic or scientific sense; i.e., it is causal explanatory theory. However, as a number have argued (e.g., Clark, 1987, 1989; Dennett, 1987, 1991c; Horgan and Woodward, 1985; Sterelny, 1990; Wilkes, 1981), the way in which 'theory' is applied to the two cases — the everyday practice of folk psychology on the one hand, and science on the other — is quite different. Indeed, *pace* Churchland, everyday psychological explanations do not count as scientific theories in any useful sense of the term. Folk psychology can certainly be regarded as *theory* (as e.g., Greenwood, 1991b, insists) albeit not theory with scientific explanatory virtues. Dennett (1991c) puts it this way: folk psychology should not be regarded as scientific causal explanatory theory, for "it does not consist of any explicit theorems or laws" (p.135). The explicit theorems and laws of nature demarcate the joints at which the world can be carved, and if folk psychology does not consist of any of them, then there is no guarantee that the kinds of folk psychology will correspond to natural kinds.

Dennett's point does not necessarily imply eliminativism, however. For even propositional attitude realism, as Fodor (1987) says, does not entail the requirement "that the folk-psychological inventory of propositional attitudes should turn out to exhaust a natural kind" (p.26). That is, some folk-psychological kinds may not pick out any (physical) natural kinds.

Like folk-psychological 'theory', many of the so-called scientific theories in psychology do not consist of explicit theorems or laws. It might be argued that these non-causal explanatory theories are therefore have no place in a developed science. I do not take this view, however; I do not wish to totally exclude these 'weaker', non-causal explanatory theories from the realm of

science (see chapter 5). Similarly, I believe that folk-psychological 'theories' need not be totally excluded from science. Indeed, if some form of the "Representational Theory of Mind" (RTM — see e.g., Fodor, 1975, 1987; Sterelny, 1990) is true, then as Fodor says, common sense is vindicated, for RTM shows how intentional states can have causal powers.

All this, however, is not meant to imply that the bold claims of the eliminativists are entirely mistaken (at least on Clark's, Dennett's, Wilkes', and my own view). As will become evident in latter sections of this thesis, we may have to submit to the eliminativist push more than many would intuitively wish to allow. ("What do you mean we don't *really* have beliefs and desires? How utterly preposterous!") Nevertheless, I take it that some species of intentionally characterized explanatory theories — possibly akin to folk psychology, but no doubt "cleaned up and made precise in various ways " (Sterelny, 1990, p.150) — are likely to survive and prosper within the scientific endeavour, for their heuristic value, if no other. If neuroscience (or, for that matter, computational psychology) provides accounts of what is actually going on in the head which contradict the intuitions of folk psychology, it is no good reason to throw the *practice* of folk psychology, *tout court*, out the window (chapter 5). For many of the terms of folk psychology may nevertheless remain extremely useful causal explanatory concepts in everyday social interaction (i.e., this suggests some form of *instrumentalism* — see Dennett, e.g., 1978a, 1987). Moreover, these terms are used to refer to the behaviours and presumed states of *whole* persons (or other organisms), and did not originate or develop as a means of identifying discrete, quantifiable states of the brain. As Sterelny (1990) correctly reasons, "even if Churchland is right in thinking that intentional psychology is badly flawed, the eliminativist moral does not follow." (p.148).

Rather than being regarded as a systematic theory of behaviour and mental life, folk psychology should be viewed primarily as a *craft* (Dennett, 1991c); "the *theory* of folk psychology is the ideology about the craft " (p.135). Folk psychology is more a form of "*bedrock* theorizing " (Clark, 1987, p.146) than a primitive speculative theory of mind. Support for this view is presented by both Clark (1987, 1989) and Dennett (1991c) when they compare folk psychology to naive or folk physics (Hayes, 1979). On this view, just as we are born with the propensity to develop abilities to assess and predict some gross behaviours of the natural world, so too are we born with the propensity to develop abilities to assess and predict some aspects of our social world. Clark (1987), for instance, argues that "just as a roughly accurate grasp of some basic *physical* principles is vital to a mobile organism, so too will some roughly accurate grasp of basic *psychological* principles be vital to a *social* organism" (p.140). By "vital" he means essential for survival, i.e., "evolutionary necessity" (p.140). (This view of folk psychology dovetails rather nicely with some theories of the evolutionary development of consciousness, which will be discussed in chapter 5).

It is important to remember that the everyday mental terms (EMTs) of folk psychology are used in social contexts to describe and explain to others (and, perhaps derivatively, ourselves) the behaviours and supposed mental states of people. As such, their full meaning can only be gleaned by studying their use in these social contexts; we must look to the social milieu to explain the nature and use of the everyday mental terms. Social constructionism (e.g., Harré, 1986) and much of recent social psychology (e.g., Fletcher, 1984, forthcoming) trade exclusively in the realms of the social milieu and the EMTs. Unfortunately an adequate examination of these theories is far beyond the scope of this thesis;

nevertheless, an argument for the inclusion of an examination of the social milieu in explaining the 'human condition', including consciousness, will emerge in the following pages. Anything approaching a full explanation of all that is entailed by 'consciousness' will require consideration of the social realm (see especially chapter 5).

That is the wide use of the term 'folk psychology'. The narrow use of the term equates it to belief-desire or intentional psychology: the explanation of a person's behaviour or mental states by appeal to the beliefs and desires of that person (agent). The distinction between the narrow and wide uses of the term is not hard and fast: essentially the narrow construal is an idealized form of folk psychology.

Beliefs and desires are taken to be states which have mental content: they refer to things apart from themselves — either imaginary or real world objects or events. So, for example, it is said that one can believe that it is snowing in the mountains during a storm, believe that there are spiders on Mars, desire a large piece of blueberry cheesecake, and believe that kiwis can fly. Thus beliefs and desires are 'intentional': they refer to or are about possible things in the world. Notice how the content of a belief or desire — in the case of beliefs, the proposition following 'that' in the clause — may be true or false independent of the truth of the statement attributing belief or desire (the propositional attitude). Kiwis cannot fly, yet one may have the belief that they can.

It is this intentional or representational aspect of the propositional attitudes, and mental states in general, which is seen as a serious obstacle to physicalism, for human action and experience is described and explained in everyday language by the use of 'intensional' terms, whereas the languages of

the physical sciences are wholly 'extensional'. It appears impossible to describe and explain human action without recourse to intensional language, yet the demands of physical science require intensional terms to be "eliminated at some stage: reduced to, or explained or paraphrased by, extensional sentences. And this is impossible: no extensional sentence, or set of such sentences, ever has the same truth-conditions as an intensional sentence" (Wilkes, 1978, p.16). While I do not claim to offer a sufficient solution to this problem here, the discussion in this and the following chapters will point the reader in the direction which I believe shows most promise: namely, that a scientific understanding of the everyday ascription of mental terms is possible, but that a reduction or elimination of these concepts by some more 'basic' science is likely to be unjustified, if not impossible. The purposes and goals of sub-personal scientific psychology differ from those of personal scientific psychology, and the purposes and goals of folk psychology differ from both[1] (as I shall make clear below).

1.2 *Purposes and goals of scientific psychology.*

It is of critical importance to distinguish the purposes and goals of scientific psychology from those of everyday or common-sense psychology (Wilkes, 1981). For whereas the latter is primarily concerned with the explanation, description and prediction of specific actions in particular contexts and circumstances, the former seeks "to identify and explain the pervasive, fundamental capacities that underlie the purposive behavior of humans and animals." (Wilkes, 1981, p.150).

---

[1] The distinctions made here owe much to Dennett's (e.g., 1978a, 1987) distinctions between sub-personal and personal psychology, and between the "design stance" and the "intentional stance".

Consequently, scientific explananda (the phenomena requiring explanation) may not correspond exactly with those states or properties identified by folk psychology, and the explanans (the terms doing the explaining) of the two enterprises will differ considerably. (The explanans of folk psychology are the propositional attitudes — beliefs, desires, thoughts, and the like — and other EMTs; they are used to describe and explain the properties and actions of ourselves and other people.) There are at least two reasons for this: Firstly, the EMTs lack the explanatory depth, precision, and consistency required of good scientific explanans (Wilkes, 1981, 1988a, 1988b). Secondly, propositional attitude talk (and, presumably, most other EMT talk) is essentially "a holistic net thrown across a body of the behavior of an embodied being acting in the world" (Clark, 1989, p.5), and hence is likely to refer to a variety of types and degrees of physiological and structural-functional states. In short, mental state kinds may not correspond to natural kinds. The tools of folk psychology are not likely to carve the psychological beast at its natural joints; the proper tools for the job are the sharp knives of good scientific theory. (Nevertheless, as Clark, 1989, points out, the terms of folk psychology do provide good starting points for scientific investigation — otherwise we would have a hard job finding much to investigate!). Granted, some folk-psychological terms are adopted by science, but if they are then they will be considerably 'tightened up' or 'adapted' (Wilkes, 1988a). What gives us reason to suppose that 'consciousness' is not a suitable term for adoption and adaption by science? An answer to this question will emerge from the discussion in sections 1.3 and 1.4, below (see also Wilkes, 1984, 1988a).

1.3 *The everyday folk meanings of 'consciousness'.*

The term 'consciousness' has had a relatively short and eventful life so far — in everyday, philosophical, and psychological usage. It first appeared in English and some other European languages in a few writings of the 17th century (Wilkes, 1988a). Not even in ancient Greek is there anything appropriately translatable as 'consciousness', or even 'mind'. Wilkes (1988a) suggests that 'consciousness' may also be somewhat idiosyncratic to post-16th century European languages. In Chinese, for example, there are no terms that adequately capture the English 'conscious(ness)'. Yet throughout this time there has been no singular, all-encompassing, universally accepted (or even widely accepted) definition or meaning given to the term. It can mean many things to many people — layperson, philosopher, and scientist alike.

Natsoulas (1978), following the *Oxford English Dictionary* (1933), discusses seven concepts of consciousness contained in everyday usage. They are consciousness as: (1) "joint or mutual knowledge"; (2) "internal knowledge or conviction": a basic knowledge of oneself and one's actions; (3) "awareness": in the general sense of being aware — of external facts and objects, that one is having a certain thought, etc; (4) "direct awareness": the ability or state of being "non-inferentially" aware of ones own thoughts and perceptions; (5) "personal unity": the up-to-date, complete set of mental episodes of a person; (6) "the normal waking state"; and (7) "double consciousness", which refers to the phenomenon evident with double or multiple personalities, where the trains of thought or mental capabilities can be viewed as independent to some degree (Natsoulas, 1978, pp.909-913).

Psychological research on folk conceptions of consciousness is scarce. One recent preliminary study by Kemp and Strongman (unpublished), however,

provides a number of interesting results concerning the definitions people give of consciousness, and how they view marginal cases (children, animals, and the retarded). Five general categories were extracted from the brief definitions of consciousness provided by the respondents in the study: sensory awareness or awareness of the environment; awareness of one's place in the environment; self awareness; awareness of the existence of others; thought, imagination, or some other cognitive ability. Thus, for the most part, consciousness is seen as synonymous with awareness. Moreover, the study found sensory awareness to be the most frequently cited definition of consciousness. Overall, however, the evidence suggests little consensus in peoples' definitions of consciousness; they did not share a coherent conception of consciousness suitable for applying to young children, animals, and the retarded (hence there was little agreement on these marginal cases).

Kemp and Strongman's study gives some support to the view that the folk notion of consciousness is varied and sometimes incoherent; people tend to be imprecise in their definitions of consciousness, and inconsistent in applying the term. If this brief study is suggestive of a general pattern amongst a larger population, then it certainly gives credence to the view that 'consciousness' is not a good explanandum for science.

Thus, in agreement with Allport (1988), Patricia S. Churchland (1988), Sloman (1991), and Wilkes (1984, 1988a), it is my contention that consciousness, at least as it is conceived in folk-psychological talk, is not a good explanandum for science because the everyday language in which it is embedded is often too vague, incoherent, and just plain mistaken about the exact nature of the physical world. It is looking increasingly likely that 'consciousness' does not denote a single entity or property unifying all the cases referred to by different

usages of the term. "[T]here is no unitary entity of 'phenomenal awareness' — no unique process or state, no *one*, coherently conceptualizable phenomenon for which there can be a single, conceptually coherent theory" (Allport, 1988, p.161). Rather, 'consciousness' denotes a complex heterogenous set of properties, events, or states; and thus it may be like the notion of understanding, which "denotes a complex set of prototypical capabilities or conditions" (Allport, 1988, p.162). Indeed, Wilkes (1978, 1984) suggests that consciousness, like intelligence (and, I would say, understanding), should be regarded as a 'second-order concept'; that is, ascriptions of consciousness will depend on the prior ascription of a range of more 'basic' first-order mental or other psychological concepts. "In other words, we presuppose a whole slew of psychological ascriptions — to do with perception, motivation, belief and desire, misperception, illusion, recognition, *etc.* — when an ascription of consciousness makes sense; conversely, where some set of first-order mental statements are appropriate, then the 'fact' of consciousness follows automatically." (Wilkes, 1984, p.238). However, there may be a crucial difference between the notions of consciousness and intelligence as second-order concepts, according to Wilkes (1984): consciousness cannot be adopted by science as an *analysandum,* for unlike intelligence, it is *too* imprecise and heterogeneous a term; it does not cover a tidy or systematically-related set of behaviours suitable for analysis. (The analytical strategy in science is discussed in more detail in section 2.3, part II of the present chapter.)

Despite the findings from the cognitive and brain sciences that suggest otherwise, I think it is still too early to say once and for all that there is really no such 'thing' as consciousness, no property or set of properties common to our multifarious subjective experience. Indeed, as will be mentioned in chapter 5,

following Dennett and Kinsbourne (1992), some form of realism about consciousness may be warranted. Nevertheless, if we are still to use the term 'conscious(ness)', and to theorize about it, we must bear in mind the points made here and elsewhere (Allport, 1988; Wilkes, 1978, 1984; and others), for not doing so will often lead us astray, creating problems where none may exist (the 'qualia' problem may well be one such case: see chapter 4). In other words, when theorizing about *consciousness* we must unburden ourselves from the shackles of folk-psychological intuitions and assumptions about its nature, function, and mystery. 'Consciousness' must be baked in a *very* hot theoretical kiln before it can become a legitimate scientific construct.

1.4 *Scientific and philosophical uses of 'consciousness'.*

The following sketch of a taxonomy (from Copeland, unpublished) provides a useful starting point for a discussion of the psychological and philosophical uses of consciousness. This taxonomy considers consciousness as : (i) the capacity to perform a set of baseline functions; (ii) a type of internal monitoring; and (iii) as sensory episodes accompanied by qualia.

(i). The *baseline sense* of consciousness.

This specifies perception of the world via sense organs, and the ability to perform such inner processes as reasoning, planning, deliberating, judging, etc. Although as Jaynes (1976, p.47) suggests, it is conceivable that there might have been a race of beings who satisfy these baseline conditions and yet we might still regard them as not having been conscious (in some further sense). Humphrey (1986) makes a similar point when he talks of "perception *sans* sensation" (p.57).

Even if such a race of beings did not exist, it is nevertheless apparent from numerous psychological studies that much of human functioning occurs without requiring conscious experience. This raises the point of whether consciousness is necessary for cognition and behaviour. Is consciousness of any use, or is it just a relatively unimportant and impotent by-product of the complex functioning of our brains? I suspect that many people would be horrified at such a proposal. Surely it must have *some* significance? Indeed, conscious experience is, to many people, the *essence* of what it is to be a normal, functioning, experiencing, and cognizing human being. The following two categories in Copeland's taxonomy of 'consciousness' should shed some more light on this issue.

(ii). Consciousness as a type of *internal monitoring*.

This proposed internal monitor (or monitoring system) allows us to be aware of some of the perceptual, cognitive, and bodily action processes, and not aware of others. A commonly given example of a process that cannot be monitored is the pupillary response. There are also processes which can be monitored, but are not monitored all the time. This internally-directed attention or awareness aspect of consciousness is limited in its processing capacity.

So we get internal monitoring theories like that of Armstrong (1968), in which consciousness is hypothesized to be "a process in which one part of the brain scans another part of the brain." (p.94). Or that proposed by Humphrey (1986), in which consciousness is viewed as a metaphorical "inner eye".

Internal monitoring models of consciousness are widespread within cognitive theories of mind. Johnson-Laird (1983, 1988a, 1988b), for example, outlines a theory of the conscious and unconscious mind based on a

computational framework. This theory postulates that 'simple consciousness' (bare awareness) can be explained by way of a high-level monitor that arises from the complex parallel processing of the brain. (Chapter 3 considers computational models of consciousness in more detail, including the modularity theses.)

All this talk of a monitoring system seems plausible, one might argue, but surely it leaves something out? The visually aesthetic pleasure of a spectacular sunset, the searing pain of an acute burn, or the sweet, rich taste of a blueberry cheesecake are sensations which appear to have special qualities that internal monitor theories cannot explain. There seems to be more to awareness than is captured in the computational-functional notions of a monitoring system. Thus we turn to the third broad construal of 'consciousness' in philosophy:

(iii). Consciousness as *sensory episodes accompanied by qualia.*

This interpretation of conscious experience is what Copeland (unpublished) calls "[t]he ineffable *feel* of it all" (p.252). Not only can we perceive objects and their sensory qualities, and be aware of them, but there is also some sort of "feel", or "subjective character", or "immediate phenomenological quality" (p.253) to this experience. *'Qualia'* (singular *'quale'*) is the collective term some people prefer to give to such special properties of conscious experience. There is something uniquely puzzling with the notion of qualia; the concept appears to defy convincing explanation. There is something it is *like* for a conscious entity to experience X (Nagel, 1974/1979), yet no amount of (current or future) third-person science seems capable of capturing these phenomenal qualities of experience. (There will be more about this baffling enigma in chapter 4.)

There have been numerous other attempts at individuating or classifying the various meanings of consciousness, some more successful than others. This in itself serves well to illustrate the looseness and generality of the term. Not only is our everyday conception of consciousness not a suitable explanandum for science, but many philosophical and scientific construals of the concept are still notoriously vague and imprecise. There is some general consensus on the broadly specified phenomena to which we normally attach the name 'conscious(ness)', for example: the awareness of sensations; the ability to plan, deliberate, judge, and the like; short-term memory; awareness of the self; and even more broadly, the different states of awareness between the waking and sleep states (see e.g., P. S. Churchland, 1988; Wilkes, 1984, 1988b). Yet there is little agreement on just how of these phenomena are to be explained, and even on whether they can be explained at all.

2. *Scientific realism and explanations of consciousness.*

2.1 *Introduction.*

"In recent decades a virtual Copernican Revolution has taken place in the philosophy of science, a radical change that has profound implications for the human sciences." (Manicas & Secord, 1983, p.399). This radical change is the realist view of science (e.g., Bhaskar, 1975; Keat and Urry, 1975; Manicas and Secord, 1983). Scientific realism is an alternative to the empiricist and paradigmatic views of science, and is now the dominant force in philosophy of science.

The task of realist science is to formulate and develop theories that in some way represent and explain the world. Hypotheses are formulated about some

possible causal mechanism(s) of structures and events in the world, usually from some patterns of experience. This leads to the construction and development of theories that detail the existence and operation of those causal mechanisms which are proposed to explain the structure or event in question. Frequently these causal mechanisms will be sets of variables or 'hidden structures' that underlie our observations of the event. Thus we get the postulation of atoms and their subatomic constituents in physics, the molecular structure of DNA in biology, black holes and cosmic strings in cosmology. These hidden variables afford our theories greater explanatory power by going beyond the observable. "We pay the epistemological price of unobservability because we value the intellectual benefits it brings us, and rightly so. We depart from strict empiricism because it leaves the world unexplained." (McGinn, 1991, p.89).

Nevertheless, unobservability is not the be-all-and-end-all of realism; nor is it likely to be a necessary central virtue of a theory. For a realist, truth is the principal virtue of a theory; the structures posited by realist theories are considered to *actually* exist. Although as Weston (1992) summarizes, truth for scientific realism may best be considered as "approximate or near truth"; truth, in this sense, is a 'horizon concept'.

A question worth asking at this point is whether consciousness — or at least some of those phenomena or properties ordinarily grouped under the term — is amenable to realist explanation in terms of some 'hidden structure(s)'. McGinn (1991) believes so: "Consciousness does have natural depth, a concealed underside. We need to extend the strategy that has worked so well in other areas to this case too: the demands of theory make the attribution of hidden structure to consciousness unavoidable." (p.91). I think this is a

reasonable claim even if we regard consciousness in the way I have been emphasizing: as a multifarious folk-psychological concept, not readily adoptable by science as a neat explanandum or explanans. However, I differ from McGinn in one important respect: he believes that consciousness has a hidden structure that will forever lie outside our explanatory grasp. McGinn thus joins Nagel (1974, 1979, 1986) as one of the "new mysterians", in Flanagan's (1990) nomenclature. I believe the new mysterians to be mistaken, however; the following section explains why.

2.2 *Are explanations of consciousness beyond our ken?*

From the point of view of our own introspection — i.e., the 'subjective feel' or 'phenomenological quality' of our conscious states ("consciousness$_p$", in Block's, 1991, terminology) — it may appear that that is all there can be to consciousness; what you see is what you get — what more could there be? But on closer examination, it is apparent that we are duped into accepting this by the vagaries and limitations of our own conscious experience, as McGinn (1989, 1991) points out. If we are naturalists about consciousness$_p$, then there must be some mechanisms and properties that explain its existence and nature (call them *P*), and these mechanisms and properties can be said to underlie consciousness$_p$, for they are inaccessible to it; we are not aware of the physical (or computational, or ... ?) goings-on that underpin our experience.

However, McGinn makes the further (somewhat extreme, but nevertheless plausible) claim that even though there is certain to exist some full explanation of consciousness (including consciousness$_p$), we humans are entirely incapable of coming to conceive or understand it. Most uncontroversially, introspection is inadequate for revealing the where and what of *P*; introspection reveals

nothing about the properties of the brain. But neither will the third-person perspective of science render *P* intelligible, as our concept-forming systems are inadequate for the job. According to McGinn (1989, 1991) we are "cognitively closed" to the explanatory concepts required by a naturalistic account of consciousness.

Certainly McGinn is likely to be correct about there being *some* limitations on our concept-forming capacities and other cognitive apparatus. Cherniak (1986) and Fodor (1983), amongst others, argue along similar lines. It is highly likely that there are limits for knowledge — that the mind is *"epistemically bounded"*, in Fodor's (1983, p.120) words — for a number of conceivable reasons. There are likely to be various quantitative limitations on the information processing capacities of our brains, for instance. Both Fodor and Cherniak pursue this point; "we are in the *finitary predicament* of having fixed finite limits on our cognitive resources" (Cherniak, 1986, p.6). Moreover, if some form of modularity is accepted, then "modular systems may be supposed to be constrained in respect of the *class of hypotheses* to which they have access, and in respect of the *body of data* that can be consulted in the evaluation of any given hypothesis" (Fodor, 1983, p.122). (Note that even if one rejects modularity for some general intelligence, it does not follow that our cognitive system is epistemically *un*bounded, as Fodor rightly argues.)

Cherniak (1986) provides a further argument for the limits for knowledge claim. This argument hinges on the theory of natural selection: natural selection 'pre-tunes' a system specifically to the given terrestrial environment, and in so doing makes trade-offs for maximum efficiency in that environment. Thus even if a system could be attuned to the vast (conceivably, unlimited) complexity and diversity of the universe — and there is little reason to suppose that this is even

a possibility — it would certainly not be good, efficient design on the part of Mother Nature to build a system that operates in this way. Indeed, I hasten to add, such a system, even if it were possible, would be such a disaster that it would hardly get off the ground, let alone have the opportunity to strive for survival; it would simply not be able to operate in the real world.

Why does McGinn believe we are cognitively closed with respect to consciousness? Briefly, it is because he believes there to be "no form of inference to the best explanation that could draw an intelligible link between any set of brain properties and consciousness" (Flanagan, 1990, p. 336). The reasoning behind this claim is that any inference to the best explanation of *P* will need to be grounded in some perceptual brain facts, and the postulation of some perceptual brain facts can entail only further brain facts, and not facts about subjective consciousness. Consciousness is not perceived by looking at the brain; nor can consciousness be explained by the postulation of perceptible brain facts, for physical concepts cannot capture subjective psychological concepts: they express completely different kinds of properties. However, as Flanagan (1990) makes clear, this "homogeneity constraint" is overly restrictive, and McGinn's argument for it "involves a relatively flat-footed trick" (p.338). We *are* permitted to draw explanatory links (inferences) between subjective reports of conscious experience (viz. sensory awareness) and brain properties, when we can establish they are reliably linked, because in such cases we have a prior commitment to the existence of conscious experience (the subject has reported it). The object of our explanatory quest is not physical phenomena alone, as McGinn would have it, but rather the reliable link between the postulated brain events and the subjective reports of experience. The former are put forward to explain the latter. (More needs to be said about the subjectivity

of conscious experience, but that awaits the discussion in chapter 4.)

Despite these persuasive arguments, I remain sceptical of any *a priori* limits on our ability to gain knowledge of *specific* features of the universe, and on our ability to understand these phenomena (and Fodor's and Cherniak's arguments do not suggest that there are such specifiable limits). I believe that McGinn is speaking too soon by setting a definite limit on the scope of scientific theorizing and our ability to understand it (the limit being consciousness); who knows what the science of some distant era might bring, let alone to what extent our concept-forming capacities might develop, and our means of expressing ideas might change. Thus we can reject McGinn's premature pessimism about the limits of our scientific understanding of consciousness without claiming that his conclusion is false; yet no amount of purely *a priori* reasoning can show that his conclusion is correct. Just because the 'problem of consciousness' appears, at present, to be so difficult, ineffable, and beyond our explanatory grasp, it is no good reason to claim that it will forever remain so. As Paul M. Churchland (1988) says, "our current bafflement does not of itself show that no neurobiological [read: scientific] understanding is forthcoming" (p.279). McGinn appears to have fallen into the trap occupied by a number of other pessimistic new mysterians (e.g., Nagel, 1974, 1979, 1986; Jackson, 1986): they view the world through 'qualia spectacles', resulting in a myopic view of physical science (I discuss the 'problem of qualia' for science in chapter 4).

2.3 *Realism and the analytical strategy.*

The central tenet of the analytical strategy is that the best explanations of psychological phenomena are typically those where the phenomena in question are treated "as manifestations of capacities that are explained by analysis" (Cummins, 1983, p.1). More generally, as Cummins (1975, 1983) persuasively

argues, and as is evidenced by its widespread use in various sciences, particularly biology, the analytical strategy is the most appropriate or correct method for explaining how dispositions are manifested in a system. Furthermore, the notion of explanatory analysis is the foundation of an important doctrine or explanatory strategy known as *homuncular functionalism*, which will be discussed in detail in the next chapter (see e.g., Lycan, 1981a, 1987).

Nomic subsumption in psychology — subsumption of psychological phenomena under causal laws — is the Received Doctrine about psychological explanation. However, nomic subsumption typically does not constitute satisfactory explanation. (Cummins, 1983.) A realist philosophy of science rejects the Humean view of causation, offering instead an account of causal explanation that requires "the discovery both of regular relations between phenomena, and of some kind of mechanism that links them." (Keat & Urry, 1975, p.30). Descriptions of the structure and operation of these underlying mechanisms are key to realist explanatory theories (see e.g., Keat & Urry, 1975; Manicas & Secord, 1983). Thus, on the realist view, the apparent regularities of events (and hence the scientific laws that we can sometimes postulate) result from the existence and operation of the causal mechanisms or properties of structures in the world, not just from the regular and contingent conjoining of these events.

The realist philosophy of science also rejects the Received Doctrine's tenet that explanations of change — the changes of state in a system — are of primary concern for psychology. The most important scientific questions are typically those concerning properties, not changes (Cummins, 1983). Research aimed at identifying the causal factors responsible for a system S acquiring a property P

will not provide substantial explanations of those properties. Rather, what is required is an account of how P is *instantiated* in S; i.e., " 'In virtue of what does S have P?' " (Cummins, 1983, p.15). Such an account is best achieved by an analysis of S. An analysis of S will postulate or identify S's components and their organization (and hence interaction). Further, the analysis must — at least eventually — appeal to the properties of those components.

Given the multifarious nature of all that comes under 'consciousness', and hence the inappropriateness of adopting 'it' as an explanandum, there seems little sense in talking of consciousness as a single property to be analyzed. Certainly the system that is held to instantiate conscious states and processes can be analyzed; and at a number of different levels, from the more-or-less purely computational to the more-or-less purely brute physical (see chapter 2); the result of this analysis being the specification of the mechanisms and processes that are proposed as being, or being responsible for, the observed conscious behaviours or the inferred conscious states and processes. But some argue that there is a further aspect of consciousness — phenomenal experience, or consciousness$_P$ — that is not amenable to third-person analysis (e.g., Nagel, 1974/1979). This matter will be taken up in chapter 4, and chapter 5 will consider the question of whether *S-F* theory is sufficient for full explanations of the phenomena of consciousness.

3. *Folk psychology, functionalism, and a compatibilist-cum-pluralist view of psychological explanation.*

The story so far: Folk psychology is first and foremost a craft or form of 'bedrock theorizing' about the behaviours and psychological states of our fellow beings, not a crude and outdated attempt at systematic, causal explanatory scientific theory. It is unlikely that many of the terms of this folk craft — the EMTs — are natural kinds: they tend not to accurately carve nature at its joints, hence they are not likely to feature as adequate or reliable explananda or explanans for science. This is no more so than for the EMT 'consciousness'.

One general and dominant strategy for explanation in science is analysis: analysis of the system in question into its constituent components, and the analysis of dispositional and non-dispositional properties of systems. The analytical strategy lies at the heart of functionalist, and particularly homuncular functionalist, theories of mind (see chapter 2). However, consciousness *per se*, understood even as a second-order concept, may not be a suitable candidate for analysis.

Given these conclusions, we are faced with a number of important questions. For example: What are the adequate explananda for science which correspond to the everyday notions of consciousness? (Some likely, but broadly specified, candidates were mentioned in part II, section 1.4, above.) Of more relevance to the present thesis is the question: What systems of scientific research allow for the best explanations of the phenomena of consciousness? The next chapter deals with the doctrine that is often claimed to be the sine qua non of psychological investigation and explanation: *functionalism*. It is an approach to studying the mind which many have claimed, or at least implied,

has exclusive dominion over explanations of the mental. At the heart of the doctrine of functionalism is the underlying urge to find a purely formal, abstract description of mind that sufficiently captures the essence of the mental.

A lot has been claimed of functionalism; some of it warranted and some of it not, for it very much depends on what version is put forward. Importantly, it has been claimed of some varieties of functionalism that they offer our best hopes of explaining the mental, including, in the more ambitious versions, overcoming the standard objections to any physicalist account of consciousness (particularly those that hinge on the 'qualia' aspect of conscious states; see chapter 4). Is some version of functionalism really our best means for explaining the mental, including that last bastion of human uniqueness, consciousness, or is functionalism a lost and dying cause? Or indeed, is some version of functionalism a necessary but not sufficient part of a general scientific approach to explaining the mental, perhaps setting us on the right track, but not in itself able to offer a complete account? The 'setting on the right track' that I refer to will become clear in the discussion of the more recent developments of functionalism, where a theoretical stance that is crucial to the present thesis will come to light — the multiple-level view of nature.

To pre-empt part of my conclusion somewhat, I think there are a number of problems with some versions of functionalism. In particular, the earlier versions of the doctrine were seriously flawed, as is now widely acknowledged; and still other versions may well be mistaken in their attempts to seek formal, in-the-head accounts of the mental states as identified by folk-psychological talk. Moreover, I take it that purely formal, abstract accounts of mind are important but not sufficient for psychological explanation. As will become clear, I take a pluralist line on psychological explanation, and find favourable the

compatibilist leanings of Noble (1990), and Sterelny (1990), amongst others. Our explanatory theories will often be at best incomplete if we continually ignore or relegate environmental and social issues. We need to combine both micro- and macro-theories in our explanations of 'the human condition', e.g., the neurobiological with the social, the individualist functional-computational with the semantical. No one approach will be sufficient on its own to establish adequate — let alone anything approaching complete — explanations. Some consider the scientific and folk-psychological views to be incompatible (e.g., P. M. Churchland, 1981). Others (the present author included) consider a reconciliation between the two to be desirable; and if it is at all possible, then it may be that *some* version(s) of functionalism will offer our best hope (Sterelny, 1990, pp.2-3). Hence the present interest in the status and nature of various versions of the doctrine.

A central question of this thesis still requires an answer, however: Is functionalism the right program for scientific psychological explanations of consciousness? In order to provide an answer, we must first explicate in some detail what functionalism is all about, and what some functionalists have said about consciousness. This is the task of the next two chapters.

Chapter summary:

Physicalism is the dominant ontology of contemporary science — There are seemingly insurmountable problems for any physicalist explanations of the mental — Consciousness poses special problems for physicalism, namely the intentional properties and subjective qualities of conscious states.

Folk psychology is not scientific causal explanatory theory — Folk psychology is a folk craft, akin to folk physics, and as such has great heuristic value in everyday life *and* (at least in some form) in science — The purposes and goals of scientific psychology are considerably different from those of folk psychology — 'Consciousness' is a multi-faceted folk-psychological concept that will not serve as a good explanandum for science — Nevertheless, given the present neonate state of the cognitive sciences, we are licensed to use the term, provided that we are very careful to dump all excess folk-psychological baggage surrounding it.

The realist philosophy of science is to be favoured over other interpretations of physicalism — Explanations of consciousness will not forever lie outside the bounds of scientific understanding — The analytical strategy is a widely used and productive means for providing good psychological explanation — The analytical strategy underlies our best versions of functionalism — Pluralism/ compatibilism is the best path towards full psychological explanation.

CHAPTER 2.

FUNCTIONALISM.

Part I. Functionalism: The wellspring of our best psychological explanation?

1. *Introducing functionalism.*

The doctrine of functionalism has been with us for over three decades now[2]. Briefly, functionalism characterizes psychological explanations of a system in terms of abstract descriptions of the components and their role(s) in the working of the system. Moreover, such characterizations of roles or functions will specify some internal states of the system — the causal interactions that obtain between a system's inputs, outputs, and other internal states. Some functionalists (whom Block, 1980b, calls "metaphysical functionalists") regard these proposed internal states as *mental* states; on this view, mental states just are functional states. It is this latter, stronger type of functionalism (what Block and Fodor, 1972, and Block, 1980b, call the "functional state identity thesis") that has received the most theoretical and critical attention, and for good reason (see below).

---

[2] Putnam (1960) is generally regarded as being the first to give an explicit exposition of a functionalist thesis. Here, and in subsequent work (e.g., Putnam, 1967), Putnam suggested a theory of mind that linked the notion of cognition as a computational phenomenon with the notion of a Turing machine ('machine functionalism'; see below).

2. *Functionalism, behaviourism, and the mind-brain identity theory.*

Functionalism was originally conceived in response to some critical problems with classical behaviourism and the 'physical state identity thesis' or 'identity theory'. Behaviourism entailed the claim that 'mental states' simply consist in specifiable behavioural dispositions; each mental state was to be uniquely identified with a specific behavioural disposition. This seems plainly false, however, when one considers that mental states often interact with each other in the production of behaviour (e.g., McGinn, 1991). Thus an account of the mental states of organisms was required that admitted of the frequently complex mediation of 'inputs' and 'outputs' that are not "wholly definable in terms of observable stimulus-response sequences" (Greenwood, 1991a, p.2).

The 'identity theory' is the thesis that individuated mental state *types* are to be identified with specific neurophysiological state *types*, i.e., the mind is the brain (see e.g., Armstrong, 1968; Place, 1962). On this view, mental state types are said to 'supervene' on physical state types; that is, to put it roughly, there can be no change in mental state without some corresponding change in physical state. However there are two major stumbling blocks for strict versions of the identity theory as a theory of mental types.

Firstly, the identity theory is too chauvinistic in its ascription of mental states. On this view, a being or other system can be a legitimate member of the mental realm only if it has a neurophysiology just like ours. This is surely a rather presumptuous and unwarranted conclusion given that we cannot rule out for certain the possibility that some alien creature or future human artifact may be a suitable and legitimate candidate for the ascription of mental states. Secondly, there is the possibility that two people can be in the same mental state, as described by our everyday talk, and yet be in quite different physical

states. In other words, the worry is that if it is conceivable for a wide variety of physical states and processes to be identified with the same mental state for different people, then there may not be any unique neurophysiological correlates to individual mental states, thus disproving the claim that mental state M is identical with brain state B.

Functionalism attempted to rectify these shortcomings by identifying mental states with abstract functional states. As Lycan (1981a) says: "We may hold onto our anti-Cartesian claim that mental state- and event-*tokens* are identical with organic state- and event-tokens in their owners, but we would be better to individuate mental types more abstractly, in terms (let us say) of the functional roles their tokens play in mediating between stimuli and responses" (p.26).

3. *The failings of early functionalism.*

In the early days of functionalism, the theory of computability had a great deal of influence on the functionalist thesis, and its implications for psychological explanation. A central concept of the theory of computability is the notion of an 'effective procedure' or 'simple machine' (Johnson-Laird, 1983). An effective procedure is a specification of a procedure to carry out the mapping specified by a computable function. The notion of an effective procedure becomes more precise when formulated in terms of a Turing machine (see e.g., Block, 1990; Johnson-Laird, 1983). Turing (1936, cited in Johnson-Laird, 1983) argued that his hypothetical machine could compute the result of any effective procedure (unfortunately this thesis cannot be proved).

The digital computer, being the paradigmatic case of a computational device, was the perfect crutch upon which computational theories of mind could rest. Functionalism's early maxim became: the mind is to the brain as a program is to the computer hardware. Just as computer programs could be formulated independently of the hardware upon which it ran, it was assumed that the mind could be studied independently of the brain. The belief was that if the mind is an information processor (which undoubtedly it is, at least in some sense), and hence a computational device, then it is a serious working hypothesis that psychological or mental states could be type-identified with Turing machine states, describable in terms of machine tables (see e.g., Putnam, 1960, 1967).

The view that formal accounts of mind could be given without recourse to matters of biology led to the 'multiple realizability' hypothesis: if a formal description of mind is possible, then the mental realm is not restricted to humans with brains. Potentially, alien creatures and robots could be mental beings, provided their internal states preserved the relations between inputs and outputs as specified by that formal description of mind.

This charitable formulation of the necessary and sufficient conditions for mental states has been parodied by numerous examples: the flow of water through an interconnected set of pipes, the simple rule-following of a person translating one set of symbols into another, and even the molecular activity in a pail of water would qualify as instantiators of mental states on this early functionalist view, as long as they preserved the relevant input-output relations. These non-standard realizations were used in an attempt to demonstrate that one or other (or both) of the two presumed defining characteristics of mental states — intrinsic qualities and intentionality (chapter 1) — could not be accounted for by computational functionalism.

One of these non-standard realization counter-examples to functionalism is known as "Hinckfuss' pail". This is how Lycan (1981a) describes this unusual case:

> "Suppose a transparent plastic pail of spring water is sitting in the sun. At the micro level, a vast seething complexity of things are going on: convection currents, frantic breeding of bacteria and other minuscule life forms, and so on. These things in turn require even more frantic activity at the molecular level to sustain them. Now is all this activity not complex enough that, simply by chance, it might realize a human program for a brief period (given suitable correlations between certain micro-events and the requisite input-output-, and state-symbols of the program)? And if so, must the functionalist not conclude that the water in the pail briefly constitutes the body of a conscious being, and has thoughts and feelings and so on? " (p.39).

An unusual case indeed, although it is by no means one that can be easily dismissed as too ridiculous to be worth worrying about. For it makes the point that virtually any physical system can be given a description, say at the molecular level, such that it can be said to instantiate one or more forms of human cognitive functioning as specified by formal functionalist theory, thus suggesting that functionalist theories of mind are vacuous.

Another counter-example to functionalism is Block's (1978) "Chinese Nation". Block used this thought experiment to illustrate the apparent absence of 'qualia' (the intrinsic qualities of sensations) from any computational-functional account of mind. Assuming the "Chinese Nation" setup is functionally isomorphic to a human, as described by the computational-functional theory, then it seems intuitively 'obvious', according to Block, that it is missing a crucial element of mentality: namely, the phenomenal qualities of experience. The "Chinese Nation" would not *feel* anything, nor would it feel like anything to be that functional system.

The "Chinese Nation" and "Hinckfuss' pail" thought experiments have been used as examples of the general objection which claims functionalism describes only the *relational* properties of mental states, i.e., the relations between inputs, internal state transitions, and outputs, and cannot capture the *intrinsic* properties of mental states (viz. the qualities of conscious experience). Moreover, the objection runs, "a state's possessing those relational properties cannot be a logically sufficient condition for its possessing those intrinsic properties (though in our world, in our case, possessing the former may be contingently sufficient for possessing the latter)" (Thornton, 1989, p.10). Although this objection was originally aimed at Turing machine functionalism, some claim it is applicable to all versions of functionalism. (This broader concern will be taken up in chapter 4, along with a discussion of the efficacy and validity of thought experiments. In section 2.3, part II of the present chapter we shall see how a much revamped version of functionalism — "teleological homuncular functionalism" — deals with the non-standard realizations.)

Block concluded that functionalism is too liberal in its ascription of mental states. If mentality depends only on a purely computational description, entirely abstracted from physical realization (e.g.,Turing machine states), then almost anything could count as a mental being. On the other hand, if only biological (neurological, neurochemical) accounts of mind are sought, then that surely unnecessarily restricts mental phenomena to the human species — i.e., mental chauvinism. Moreover, it soon became evident that the early functionalist theories fell into the same trap as the physical-state identity theories: two people could be in the same mental state and yet be in different computational states. (The converse of this — that two people could be in the same computational state and yet be in different mental states — forms the crux

of the 'inverted-qualia' objection to functionalism; see chapter 4.)

If functionalism is to survive and prosper, its theories must strive for a balance such that our ascriptions of mental states will not be limited solely to humans, and yet not be so generous as to include such systems as the "Chinese Nation" or "Hinckfuss' pail". Nevertheless, since humans are the paradigmatic case of mental beings, maybe our theories of mind should take a closer look at the brain. When our goal is to seek the 'essence' of the mental (if indeed there is such a thing), then perhaps a study of the brain as a computational and representational device will reveal, or at least point to, the necessary and sufficient conditions for mental states.

Part II of the present chapter goes on to review the newer versions of functionalism that supposedly overcome many of the above mentioned problems with earlier versions. It will be seen that early functionalism's separation of function from structure, based on the software-hardware distinction in computers, is seriously misleading. A strict dissociation of function from structure is not appropriate for theories of the mind-brain. (Indeed, the division between software and hardware is not always appropriate in the case of computers.) In particular, it will be seen that function and structure are inextricably linked. There are two main points to note here: (i) the notion of function in functionalist theory denotes more than simply an abstract mathematical formulation tied to the theory of computation; and (ii) the notion of structure need not (and indeed, in functionalist theory, does not) make specific reference to the actual physical *stuff* that constitutes the system under investigation.

Part II. Functionalism Revised: Homuncular Functionalism and Microfunctionalism.

1. *Introduction.*

The idea that the human cognitive system could be modelled by a set of Turing machine states did not last long, for reasons such as those outlined above (see also e.g., Lycan, 1979, 1981a, 1987). Nevertheless, it is fair to say that the legacy of this approach — that the mind-brain is a computational device — forms the backbone of the cognitive sciences today. The thesis that the human cognitive system is a computational device, at least in some sense of the term, lies at the heart of any version of functionalism (albeit sometimes implicitly). As Johnson-Laird (1983) puts it, "if both Turing's thesis and functionalism are correct, any future theory of the mind will be completely expressible within computational terms." (p.10).

This part of the present chapter outlines what seem to be the best versions of the doctrine currently on offer — *homuncular functionalism*[3] and *microfunctionalism*[4]. A point in favour of homuncular functionalism and possibly microfunctionalism is the claim that they avoid many of the pitfalls of Turing machine functionalism, i.e., the standard functional-state identity thesis.

---

[3] Dennett (e.g.,1975, 1978a, 1978b) and Lycan (1981a, 1981b,1987, 1990) are the main advocates of homuncular functionalism, having both developed very similar theses, albeit apparently separately. However, Dennett doesn't use the term as often as Lycan; indeed Lycan appears to have coined the term 'homuncular functionalism' and its conglomerate, 'homunctionalism' (c.f Dennett's, 1978a, comment: "about such features I am a straightforward type-intentionalist or 'homuncular functionalist', as Lycan calls me ", p.xx).

[4] Clark (1989) is responsible for the term "microfunctionalism", which is used to refer to formal, abstract accounts of mind grounded in the neurally-inspired connectionist/ PDP theories, i.e., located at a finer-grained level (or better, levels) than traditional symbol-processing, sentential formulations of functionalism (traditional formulations that are concerned with the level of a "semantically transparent system"; Clark, 1989, p.35).

2. *Homuncular Functionalism..*

2.1 *What is homuncular functionalism?*

Homuncular functionalism, according to Sterelny (1989, 1990), takes the key element of functionalism — a description of the functional role of the internal (mental) states of an intelligent system (the "what it does, not what it is", 1990, p.13) — and marries it with the view that the mind has a modular[5] architecture, then applies these two notions recursively. The crucial steps here are (1) the functional-role analysis of complex, intelligent systems into less complex, less intelligent, interacting subsystems or modules (homunculi); and (2) the recursive application of this analysis, such that the proposed homunculi are progressively more simple or 'stupid', to the point where their functional roles are so specialized and simple that they can be occupied by primitive, mechanistic processes (i.e., are "psychologically primitive"; Sterelny, 1990, p.13). Thus we see the analytical strategy (chapter 1) as the central theme underlying homunctionalism.

Some ideas central to homuncular functionalism can be found in the work of writers not usually associated with the position. For example, Wilkes (1978) regards functionalism as resulting in "a descending series of functional analyses nestling into one another like Chinese boxes, breaking down the complex structures of the brain into smaller and more specialized functions and structures." (p.64). She thus favours a brand of functionalism that is somewhat closely tied to physical structure (a rather significant issue, to which I will return below).

---

[5] I am here using 'modular' to mean any functionally distinct cognitive component or 'homunculus', not Fodor's (1983) more limited notion of modularity (see chapter 3).

Dennett (1975, 1978a) discusses the strategy in artificial intelligence (AI) which parallels that of homuncular functionalism (indeed, in Dennett's view, the strategy originated in AI research):

> "The AI researcher starts with an intentionally characterized problem (e.g., how can I get a computer to understand questions of English?), breaks it down into sub-problems that are also intentionally characterized (e.g., how do I get the computer to recognize questions, distinguish subjects from predicates, ignore irrelevant parsings?) and then breaks these problems down still further until finally he reaches problem or task descriptions that are obviously mechanistic" (1978a, p.80).

Thus we see that this strategy is very much a 'top-down' approach, and that the characterizations of the functional parts ('black boxes') are inextricably intentional.

Minsky (1985), a prominent AI researcher, develops his own homunctionalist perspective, in which he characterizes the mind as a "society" of mindless agents. "How can intelligence emerge from nonintelligence? To answer that, we'll show that you can build a mind from many little parts, each mindless by itself." (p. 17).

All these writers recognize the dreaded problem that any theory of mind must eventually confront and solve: the problem of undischarged or intelligent homunculi. Our explanations of intelligence and other psychological phenomena must not reintroduce the very thing which they are intended to explain (the so-called 'Rylean regress', after Ryle, 1949). Homuncular functionalism can be shown to avoid this problem (see e.g., Lycan, 1987).

On just this point, Fodor (1968), in one of the earliest explications of a homunctionalist perspective, has this to say:

> "We refine a psychological theory by replacing global little men by less global little men, each of whom has fewer unanalyzed behaviors to perform than did his predecessors. Though it may look as though proceeding in this way invites the proliferation of little men ad infinitum, this appearance is misleading.
>
> A completed psychological theory must provide systems of instruction to account for the forms of behavior available to the organism, and it must do so in a way that makes reference to no unanalyzed psychological processes. One way of clarifying the latter requirement is the following. Assume that there exists a class of elementary instructions which the nervous system is specifically wired to execute. Each elementary instruction specifies an elementary operation, and an elementary operation is one which the normal nervous system can perform but of which it cannot perform a proper part." (p.629).

Thus Fodor describes the connection between the analytical strategy and computational-functional characterization.

2.2 *Homunctionalism's promise.*

A pertinent question to ask at this point is: What is the motivation for a homunctionalist perspective; i.e., what are the reasons for pursuing a homunctionalist study of the mind? We have already unearthed the gross underlying motivations of functionalism as a whole (part I of the present chapter), and one of the central underlying aims of homunctionalism is to do justice to these motivations. I concentrate on the homunctionalist thesis here, for it is currently the most explicit and viable version of the doctrine. Nevertheless, most of what I have to say in the following sections is also relevant to the discussion of microfunctionalism. Microfunctionalism is less explicitly a functionalist thesis, in the traditional sense, i.e., amongst other things, it is not so directly tied to — indeed, need not be tied at all to — the 'Representational Theory of Mind', in its computational 'language of thought' forms (see e.g., Fodor, 1975; Sterelny, 1990). Indeed, Clark (1989) considers the possible

inappropriateness of including this type of account of the mind as a species of functionalism. However, on a wide interpretation of functionalism at least, micro*functionalism* is an appropriate name, for it is a project that seeks the essence of the mental in terms of "patterns of nonphysically specified internal state transitions suitable for mediating an input-output profile in a certain general kind of way " and thus it "in effect identifies functionalism with the claim that structure, not the stuff, counts." (Clark, 1989, p.36).

A lot is claimed of homuncular functionalism, especially its facility to avoid the major pitfalls of the earlier versions of functionalism, and in particular, of machine functionalism (see e.g., Lycan, 1979). For instance, Lycan (1981a, 1981b, 1987) claims homunctionalism to be the prime candidate to achieve a balance between Block's (1978) charges of excess liberalism or chauvinism (part I, above). Further, Lycan (1981a, 1987) claims that homunctionalism is capable of meeting the various counter-example objections to functionalism, including the qualia-based ones.

Block's (1978) chauvinism/ liberalism criticisms were aimed specifically at the detail/ generality of functional characterizations of inputs, outputs, and internal states, and thus the serious functionalist owes Block a detailed account of functional characterizations that avoid or overcome those objections. There are two significant suggestions in the literature on this issue: a more general attack by Kitcher (1985), which will be discussed in section 3, and Lycan's (1981a, 1981b) in-depth reformulation of the functionalist thesis.

Lycan's (1981a, 1981b) suggestion for where to look for a solution to this "problem of inputs and outputs" (Block, 1978) is a homuncular version of some sentential (language of thought) account of mental representation (hence of belief, thought, etc.) An explication and critique of Lycan's position here is

beyond the purview of this thesis; suffice it to say, however, that such claims are contentious. It may be that this type of solution is currently the best on offer, as Fodor (1975) contends. However, there are other accounts of cognitive functioning, and hence mental representation, that, some argue, offer advantages over, and are better approximations to the truth than, conventional symbol-processing, sentential accounts. Some form of microfunctionalism is the major contender here, i.e., the "brain's eye view" rather than the "mind's eye view" (Clark, 1989; see section 4). Nevertheless, if Lycan (1981b, 1987), Sterelny (1990), and others are at least partially on the right track, some sentential homuncular account is likely to provide at least a major path through the quagmire of problems posed by the notion of mental representation, its intentional nature, and functional characterization. Indeed Clark (1989) himself admits that a microfunctional account need not exclude the possibility of some sentential-computational account also featuring in our best theories of the mind. It is also plausible to suppose that our best sentential homuncular accounts may be very different from the 'received view', i.e., the 'Field-Fodor-Lycan theory' (Sterelny, 1990), even though the received view has a lot going for it, including allowing "propositional attitude psychology to be integrated within theoretical cognitive psychology" (Sterelny, 1990, p.142).

To summarize thus far: Early and so-called standard versions of functionalism come up against a number of now standard objections (particularly those allied in Block and Fodor, 1972; Block, 1978). Can functionalism save itself from a premature demise by meeting or avoiding these objections? The brightest star so far to have appeared on the stage is homuncular functionalism — a case of the "analytical strategy" (Cummins, 1975, 1983) — teleologically construed (Lycan, 1981a, 1981b, 1987).

2.3 *Teleological functionalism and the multiple-level view of nature.*

Cummins' (1983) interpretation and use of the concept of function in functional analysis seems somewhat circumscribed or impoverished. Millikan (1989) hits the nail on the head when she says that Cummins, amongst others, construes function as referring "only to current properties, relations, dispositions or capacities of a thing" (p.292). What is important for Cummins is that an item has a function if it functions *as* something (e.g., as a face recognizer, a face feature detector, a zero-crossing detector). He ignores or waves aside all notions of *historical purpose* in his talk of function.

Millikan (1989) describes an historical notion of "proper function", where function is inextricably linked to 'intentional use and design' construals of purpose, or at least *as if* the intentions and purposes of designers and users were operating (the prime candidate in this latter case being evolution via natural selection; see e.g., Lycan, 1987; Sober, 1990; and below). Moreover, she claims that explanatory theories will need to make productive use of the notion of "proper function". Teleological talk of purposes and "proper functions" of organisms and their subsystems is to be encouraged, when teleology is interpreted biologically, explicated in evolutionary terms. (Lycan, 1981a, 1987; however, he leaves the detailing of such an evolutionary construal of teleology to the biological philosophers — see his 1987, p.43. See also Millikan, 1989; Sterelny, 1990.)

Some aspects of the process of natural selection are remarkably similar to the characteristics of conscious (purposeful or intentional) design. Complex organisms have the design and construction they do — that is, the intricately interrelated complex of organs, other subsystems, and their sub ... subsystems — because of the selection pressures created by the environment during the

evolutionary history of those organisms. The subsystems selected for, and their internal organization, are just so because they either have (or had) specific tasks to carry out in the promotion of survival of the individual, or 'rode in on the back of ' another component or trait that was selected for. (The distinction here is between selection *for* and selection *of* a subsystem or trait: Sober, 1984, cited in Sterelny, 1990. See Sterelny, 1990, for mention of further stipulations on the acceptance of teleofunctionalism).

Functional concepts are teleological concepts — a concept that is no less applicable to explanations of the mind-brain than it is to the liver or heart. How then, does homunctionalism fit into this picture of teleofunctionalism? Firstly, as Sterelny (1989, 1990) points out, the homunctionalist's modular view of mind-brain is certainly evolutionarily plausible. Intelligence did not suddenly emerge in a single step as some property or other of a simple unintelligent system; it is far too complex a set of phenomena for that. Nor can we suppose that increases in intelligence are just a matter of increases in the number of neurons — *organi*zation matters. For instance, without a modular architecture, it certainly would have been much more difficult for new capacities to have been added to systems, and old ones extended (Sterelny, 1990).

Secondly, teleological homunctionalism must be developed within (and indeed it promotes) a multi-level view of nature (Lycan, 1987) — a view that implies the existence of multiple levels of psychological description and explanation. Wimsatt (1976), from whom Lycan appears to draw much of his inspiration, has this (amongst other things) to say about levels of nature:

> "I will assume that being at a given level is a property primarily of things in the world: phenomena, objects, properties, processes, causes and effects, etc., and derivatively of linguistic things relating to them: descriptions, law-statements, theories, predicates, etc. Intuitively, one thing is at a higher level than something else if things of the first type are composed of things of the second type, and at the same level with those things it interacts most strongly and frequently with or is capable of replacing in a variety of causal contexts." (p.215).

In giving an account of the multiple-level view, Wimsatt (1976) points to yet another hang-over from the empiricist-positivist tradition in psychology. There has been a distinct lack of recognition in much of psychology of the idea that neither function or structure take precedence in science; both are roughly equally important, and each affects the other. The failure to recognize this important feature of the scientific enterprise continues in many quarters of psychology, despite the now fairly widespread belief to the contrary in many other sciences, following the debate over this issue in 19th-century biology.

As a consequence of the multiple-level view, functionalist theories are no longer to be fleshed out solely in terms of *either* formal descriptions, abstracted from specific cases of instantiation, *or* present dispositions or capacities of biological systems, but rather in terms of both. The function-structure (software-hardware, role-occupant, etc.) dichotomy is a misnomer. Any function-versus-structure distinction is very much a rule of thumb and heuristic distinction, rather than an ontological distinction. Functional and structural properties exist together at almost any level one cares to look at, e.g., from at least the level of individual cells, through the levels of whole organs and organisms, to possibly the levels of social organizations, and beyond. A better name for functionalism may be "structural-functional (*S-F*) theory" (Wilkes, 1981).

While on the issue of the nature of the terms 'function' and 'structure', it will be worthwhile to mention Ramsey's (1989) clarification of the meanings of them as used in science. For there seems to have been a great deal of confusion surrounding the use and construal of this distinction, particularly in relation to the computer model of the mind and its apparent corresponding distinction between software and hardware. Software is generally used to refer to the *program* that a computer executes, but this term is ambiguous. A program can refer to either the causally inert, descriptive flowchart or algorithm, or "to a list of commands in a computer language which actually *causes* the machine to function in a certain fashion" (Ramsey, 1989, p.141). It is the latter sense of program that is referred to by the term software. The sense of program appealed to in functionalism — for example, in the claim that the mind is to the brain as a program is to a computer — is only that of the former interpretation. That is, if the mind can be likened to a program, then it will only be so in the sense of a program as a descriptive flowchart. Notice that this is a relatively weak and innocuous claim, but one that underlies virtually the whole of the functionalist methodology (except maybe some connectionist versions).

What bearing does the ambiguity surrounding the notion of program in the software-hardware distinction have on the function-structure distinction? It is that the conventional software-hardware distinction *does not* correspond to the function-structure ('functional state-physical state') distinction. For "[t]he former is the distinction between a set of explicit procedural rules and the machine which follows these rules", whereas "[t]he latter is the distinction between different ways of describing and classifying various processes, by focusing either on physical properties or the more abstract functional properties." (Ramsey, 1989, p.141).

There is also a problem with interpretations of 'hardware': it can be used to refer to the actual *stuff* out of which the entity is constructed (the most common interpretation), or to the organization or configuration of the computational architecture. As will be obvious by now, the functionalist is generally not concerned with the specifics of physical implementation (indeed, it is this which is considered to be interchangeable, i.e., the thesis of multiple instantiability), and so it is the latter construal of hardware that corresponds to structure in the function-structure distinction.

To repeat the point, *S-F* theory need not be tied solely to biological entities, nor does it need to specify particular physical organs or other systems. To speak of structure is to speak of the organization of the parts of the whole, their connections and interactions. It does not necessarily imply the inclusion of the details of the specific 'stuff' out of which it is constructed. Unless, that is, those descriptions specify properties of the stuff that are proposed to have bearing on the capacities of the system or other phenomena in question (e.g., as the connectionist or microfunctional theories contend; see section 4). However, even these accounts of the properties relevant to the phenomena in question are likely to be abstracted as much as possible from the physical stuff of implementation, even if it is only to avoid charges of excess chauvinism.

2.4 *Some consequences of homunctionalism and the multiple-level view.*

Nature is multi-levelled, and its explanations and descriptions will be correspondingly so; attempts to neatly carve up explanatory theories of the natural world into the purely abstract functional and the purely concrete structural are misguided (see especially Wilkes, e.g., 1978, 1981). This is not to say that neither approach is explanatorily useful; it's just that neither on its own will provide anything like full or 'complete' explanations (see Lycan's, 1987,

discussion of this point — i.e., "the teleologicalness of characterizations is a matter of degree", p.43). This is particularly the case with many psychological phenomena, especially those that carry the 'mental' tag (i.e., those folk psychological or EMT states for which identifications were sought by the neurological and functional identity theorists).

Accepting the multiple-level view of nature forces some interesting and rather significant implications. For instance, it suggests reasons for the machine functionalists' misguided attempts to identify mental states with Turing machine states (i.e., identifications of the sort: 'to be in mental state M is for one's cognitive system to realize the machine table state(s) T_1 (T_2, T_3,...)'). We can now say exactly what the major problems are with this and other similar attempts at type-identifying mental states: (1) The to-be-identified mental states are typically those of everyday or folk psychology — a higher-level, 'intentionally infected' description, unconcerned with the more 'base level', underlying features of our mental lives. Thus, given the likely incompatibility of most folk-theoretical terms with scientific postulates (chapter 1): (2) Our best attempts at solely functionalist theorizing are unlikely to yield specific or strict isomorphisms between the mental states of folk psychology and the states posited by a mathematical theory that disregards any notion of structural organization, let alone the physical stuff.

The problems here are with both aspects of the functional-state-identity project: (1) Certainly any cognitive system will be in the midst of highly complex and intricate, ongoing activity when we describe them as being in certain mental states, but this activity will need to be specified both functionally *and* structurally, at all manner of different levels. (2) While it may be legitimate practice to make use of the everyday mental state terms in some scientific

endeavours, there is no guarantee that they pick out just those capacities postulated and identified by the cognitive and neuro-sciences. "If all this is indeed so ... then it is evidently impossible to regard the referents of ordinary terms as states of which a *S-F* theory might be true or false." (Wilkes, 1981, p.153).

The teleological homunctionalist view also counters Searle's (1980, 1984) monistic 'emergentism', in which the mental is a set of "surface properties" (the macrostructure) arising from the operations or "causal powers" of the brain (the microstructure). Searle offers this view in opposition to the non-Cartesian dualist view of there being some 'gap' to fill between the brain and the mind (Searle's stalking horse being "strong AI"). From the perspective of the multiple-level view, Searle is certainly on the right track when he speaks of micro- and macrostructure, but he is wrong in limiting its scope or application. Micro- and macrostructure are not levels of description set in concrete, but are relative terms. Moreover, from the perspective of the multiple-level view, the only 'gap' between the mind and the brain is that between the world as we know it from our own experience and the world as it is independent of any one particular viewpoint; there is no gap in the way the world is (say, between the physical and the mental). Computational psychology may have been devised as a 'gap-filler' or bridge between what we know of the brain from the third-person perspective of science and what we 'know' of the mind from the first-person perspective of experience, but it is not presumed to identify a distinct ontological level on a par with the neuroscientific brain and the first-person mind. For as the multiple-level view of nature reminds us, there is no single ontological level of the brain; rather, structural and computational-functional properties of the brain exist at all manner of levels. Furthermore, it is possible

that these descriptions of the structural and computational-functional properties of the brain may encroach on the provenance of the first-person mind — i.e., it is possible that *S-F* theory may eventually explain the first-person perspective. (Chapters 4 and 5 will further discuss the gap in our understanding of how the third-person brain relates to the first-person mind.)

If computational psychology is open to attack, it must be on different grounds than those of any 'gap-filling' interpretations.[6] Connectionist theories have so far provided the most concerted of such attacks, i.e., connectionism is seen by some to refute many of the claims made in the service of traditional computational psychology. Nevertheless, there is now a growing body of support for the position that connectionism is not necessarily a mutually exclusive alternative to more traditional computational theories. (For discussion of this issue, see e.g., Clark, 1989; Sterelny, 1990.)

One notable line of attack on computational functionalism are the non-standard counter-example critiques. Let us see why Lycan does not consider these objections to be a threat to teleological homunctionalism. Take, for example, "Hinckfuss' pail": the molecular activity in a pail of water could conceivably briefly instantiate a human program. Lycan's (1981a) reply to this unusual counter-example objection to functionalism is twofold. Firstly, the sense in which "realize" is used in the pail of water example is not the same as a revised (i.e., homuncular) functionalist construal of the term. The pail of water can be said to "realize" a functionalist program in some basic "mathematical sense, in which 'function' is synonymous with 'mapping' " (Lycan, 1981a, p.27).

---

[6] Indeed Searle (e.g.,1980, 1984) provides some valiant attempts at such arguments – most notably his famous "Chinese Room" parable – although again these can be shown to be faulty; see e.g., Block (1990); Copeland (unpublished); Hofstadter and Dennett (1981); Sterelny (1990).

However, this sense of functional realization is too impoverished, for such realization is too easily achieved, leading to ascriptions of mentality that are too liberal. The mathematical realization of a human program by a physical system such as a collection of H_2O molecules is merely fortuitous.

The second prong of Lycan's reply to the pail of water counter-example is that the mathematical notion of functional realization needs to be replaced by some more robust account that incorporates the notion of "functional *organization*, or organic integrity and autonomy" (Lycan, 1987, p.33). The notion of 'function' in functionalist theory will then be seen not simply as a mathematical 'mapping', but rather as a description of how an entity functions *as* an entity with a *purpose* or function within and *for* a system, relative to what that system is doing. Moreover — and this is a point which Lycan is not too clear in making, as Millikan (1989) points out — the notion of function should be understood within an historical context, i.e., as "proper function". Thus functionalism, properly construed, is teleological; the pail of water characterization of functionalism is not teleological. The appeal to functional organization indicates that the inputs, internal states, and outputs of a system worthy of consideration as a species of mental being are organized, related, and structured according to a *'prearranged plan'*. That is, the internal processors of such a system must be of the right type and structure such that they are capable of fulfilling the function or purpose for which they were 'designed', and they must be "organized in the relevant way" (Lycan, 1981a, p.41). Hinckfuss' pail of water does not meet these requirements "precisely because it is not organized in the relevant way, even if the de facto motions of some of the molecules in the pail happen to ape the motions that would be made by an organism that *was* functionally organized on the human model." (Lycan, 1981a, p.41).

Churchland and Churchland (1981) make essentially the same point in

discussing Block's (1978) "Chinese Nation" non-standard realization counter-example. The problem with this and other similar counter-examples to computational functionalism is that they only indict the view that mentality is dependent purely on the preservation of identical input-output relations (i.e., "Turing equivalence"). But that is not an indictment of computational functionalism *tout court*, for all the computational functionalist needs to assert is that what is minimally required for a system to be ascribed mental states is for it to be "computationally equivalent to us" (Churchland and Churchland, 1981, p.134). In other words, the internal structural-functional organization must be equivalent to ours at the (as yet unspecified) relevant level(s); "it must have a system of inner states whose causal interconnections mirror those in our own case." (Churchland and Churchland, 1981, p.134). Churchland and Churchland's (1981) reasoning for this conclusion is based on the incredible computational complexity of the human brain. Because an extraordinarily large number of computational states can occur in the brain, due to the enormous number of neurons and an even greater number of possible interconnections, it would simply be impossible to construct a Turing machine (or a Turing-equivalent machine like the "Chinese Nation") that could even approach computational equivalence; "*no* brute-force one-device/ one-square realization of a Turing machine constructible in this universe could even begin to simulate your input-output organization." (p.135).

In conclusion, functionalism properly construed is not restricted to purely formal, mathematical descriptions of a system, for such an interpretation of functionalism ignores all notion of teleology and "proper function". But more importantly, as the multiple-level theory reminds us, function and structure exist together, and thus a *purely* abstract, formal (mathematical-functional)

account of a system is not sufficient for explaining the behaviour of that system at any level; not even at the level of primitive processors — a point that many functionalists have failed to grasp (especially the machine functionalists).

3. *In defence of formal, abstract accounts of mind: the continuing search.*

Functionalism can be seen as a search for the 'essence' of the mental. Homunctionalism in particular stays true to the realist's quest for "the 'nature', 'essence', or 'inner constitution' " (Keat & Urry, 1975, p.30) of the entities or properties in question — by describing the nature and operation of the underlying mechanisms. Yet it is not clear whether this or any other type of psychological explanation currently on offer "is identical with the project of seeking the essence of the mental" (Clark, 1989, p.22); where the latter is understood "as the search for the necessary and sufficient conditions for being in some mental state" (Clark, 1989, p.22). Or more correctly, given the points made in this and the previous chapter, the problem is that the formal descriptions of *any* version of functionalism may "never isolate a class of physical mechanisms capable of supporting the rich, flexible *actual and counterfactual* behavior that warrants ascribing mental states to the system instantiating such mechanisms." (Clark, 1989, p.178, original emphasis).

I believe that the above defence of *S-F* theory against the non-standard counter-example objections to be on the right track. Furthermore, in line with the points made in chapter 1, we could do well to recognize that the mental states we commonly ascribe are everyday social constructs of the linguistic community, i.e., they are 'everyday mental terms', and are not, or at least are not likely to remain, *scientific* constructs, as Wilkes (e.g., 1981) is at pains to

emphasize. Thus, the decision to ascribe the mental states *of everyday folk psychology* to non-standard cases lies only with the linguistic community. If, however, scientific theories establish the essence of the mental — which, to repeat, is understood as the isolation of "a class of physical mechanisms capable of supporting the rich, flexible *actual and counterfactual* behavior that warrants ascribing mental states to the system instantiating such mechanisms." (Clark, 1989, p.178) — then we will have more rigorous grounds for accepting or rejecting the non-standard cases from the realm of the mental; i.e., here the decision will rest with the scientific community. And the sceptic has no legitimate right to claim that the non-standard cases provide good *a priori* reasons for the inadequacy or vacuousness of *S-F* theory in establishing the essence of the mental, for that is an empirical question. Furthermore, at present a major part of *S-F* theory rests on the theory of computation (the notion of primitive processors and what they do), and as Block (1990) points out, "[i]t is an open empirical question whether or not the computer model of the mind is correct." (p.261).

Thus we can say that the *explanatory value* of posited functional and computational kinds will save them from the criticism of vacuousness supplied by non-standard counter-examples. Objections to functionalism such as Hinckfuss' pail will hold water only if the posited functional and computational kinds do not have any significant explanatory value, in which case we would indeed "get hyperinflation of computational theory" (Sterelny, 1990, p.206); but that is simply poor *S-F* theory, and not a criticism of the doctrine with any bite.

Thus I agree with Kitcher (1985) when she argues that

> "[i]n general, Psychofunctionalisms will be vulnerable to the charge of chauvinism only if they make falling within the purview of a particular theory or a particular type of theory a necessary and sufficient condition for being a psychological entity. Then the genuine possibility of unexpected cases and indeterminate cases makes the enterprise look dubious. However, there is no reason for a Psychofunctionalist to buy into this project. It is perfectly reasonable for a Wide Functionalist, for example, to develop and use his classifications of psychological states without haboring the belief that he has hit upon the one and only true mark of the mental " (p.94).

Kitcher's argument rests on the claim that cognitive psychology has been held "to taxonomic standards that other sciences routinely rise above" (1985, p.78); namely, that the robust theoretical constructs of a science need not be, and indeed most often are not, precisely defined in terms of necessary and sufficient conditions. ("It is a bad day for 'realism' when only phenomena with hard-edged necessary and sufficient conditions can be considered real." Dennett and Kinsbourne, 1992, p.239.) Thus, even though it may be impossible to define 'mental' or 'psychological', we cannot as a consequence rightly claim, as Block (1978) does, that functionalism paints itself into a dead-end corner from which it may never successfully emerge (viz. 'the problem of inputs and outputs' or the chauvinism-liberalism problem, as illustrated by Block's non-standard or indeterminate cases).

The above point not withstanding, Kitcher (1985) admits that there are likely to be many insurmountable problems with bad versions of functionalist computational psychology. Nevertheless, it cannot be shown to be so for the reasons that Block (1978) puts forward. "The point I have been urging against Block is that computational psychology can offer theories of great scope and power without denying the actual or potential diversity of nature. For an approach can be the great unifying idea for a large range of phenomena,

without claiming absolutely exclusive dominion. Computational psychology has nothing to fear from unexpected or indeterminate cases." (Kitcher, 1985, p.96).

Now that we have cleared the way for some form of *S-F* theory to proceed, unhindered by indeterminate counter-examples, it is time to come clean on my defence of formal accounts of mind (if it is not already obvious — see chapter 1). While being an advocate of the formal approach to studying the mind (viz. structural-functional and/ or microfunctional theory), I reject the traditional computational and functional accounts. For these traditional accounts, exemplified by "GOFAI" (good old-fashioned AI: Haugeland, 1985), attempt to model or create intelligent, mental beings by putting descriptions of the folk psychological states back into their heads. As we have established, however (chapter 1), the states posited by folk psychology are not precise descriptions of the actual goings-on inside the heads of intelligent creatures. The everyday mental terms constitute a level of description of persons or systems as a whole, as Dennett (e.g., 1978a, 1987) often reminds us. EMTs are not likely to be accurate pointers to what is actually going on in the head. Consequently, a proper science of the mind will need to abandon the approach of "putting tokens of ordinary contentful talk back into the head (classical cognitivism)" and should instead seek "an account of how what *is* in the head enables the holistic ascription of such contents to the subject in the setting of the external world." (Clark, 1989, p.59).

Thus a new way of viewing functionalism (*S-F* theory) is beginning to emerge: functionalism should be considered not as the attempt to identify functional state types with mental state types (i.e., the functional state identity theory), but rather as the attempt to map the structural-functional (cognitive)

architecture of the brain. *S-F* theory thus encompasses what Dennett (e.g., 1978a, 1987) calls the design and physical stances, as opposed to the "intentional stance", which embodies the ascription of mental states.

Abandoning the more traditional interpretations of computational functionalism does not obligate one to forsake the conception of the mind-brain as a computational device, however. It certainly should be possible, and indeed desirable, to retain some formal account of what is actually going on in the head; and given the discussion of the preceding sections, there are likely to be a number of different possible types of such accounts, collected together under the general title of *S-F* theory. We turn now to one such candidate which, as one of its main proponents claims (Clark, 1989), may have proprietary rights over the territory: microfunctionalism.

4. *Microfunctionalism.*

4.1 *What is microfunctionalism?*

Connectionism or PDP theory is, for the most part, seen as an alternative to the traditional sentential-computational theories. These new theories of cognitive functioning can be interpreted as complete alternatives to the traditional accounts, but more recently some authors have given serious consideration to some possible non-mutually exclusive accounts; still, the debate continues. (For general discussions of connectionism, see e.g., P. M. Churchland, 1988; Clark, 1989; McClelland and Rumelhart, 1986; Smolensky, 1988; Sterelny, 1990, amongst others.)

A substantial part of Clark's (1989) book is an exploration of the kind of formal description (i.e., abstracted as much as is possible from particular

physical realization) that he sees to be the best on offer — a "microfunctional" description, based on the general PDP/ connectionist approach. Adding support to Clark's (1989) claim that connectionism can rightly be branded a species of functionalism, is Ramsey's (1989) argument, *pace* Thagard (1986), that functionalism is not necessarily tied to traditional computational accounts of cognition (or at least the serial architecture that it is based on). Functionalism has nothing to fear from connectionism. Connectionism and functionalism will appear incompatible only when the notion of structure or hardware is misconstrued as referring to the actual physical stuff rather than the *"computational architecture* or *causal arrangement* of such stuff " (Ramsey, 1989, p.143, original emphasis). This is not to say, however, that as good functionalists we should completely ignore neuroscience, as Ramsey (1989), Thagard (1986), and others are quick to point out. Neuroscience is likely to provide useful insights into the functional nature of cognitive states.

Connectionist models tend to posit explanations of the emergence of higher levels of a phenomenon from lower levels in terms of statistical or probabilistic functions, whereas homunctionalism takes a more mechanistic approach, decomposing a complex system into simple, interacting parts. Connectionist theory suggests that the discrete, 'digital' primitive processors posited by traditional computational functionalism (exemplified in Turing machine states) may not be the underlying mode of cognitive operation. Instead, connectionism suggests what is claimed to be a more neurally plausible alternative: the primitive processors of the brain are more 'analogue' or continuously variable in their operation. Nevertheless, connectionism does not preclude discrete input-output functions at higher levels of description.

Thus the basis of computational functionalism is undercut by connectionism, hence connectionism appears to threaten the analytical strategy

of homunctionalism (Bechtel, 1991). Indeed it is at this point where Clark (1989) diverges from Lycan's account of functionalism, i.e.,*homuncular* functionalism.[7] Functionalism as a general doctrine may not be under threat from connectionism, as Clark (1989) and Ramsey (1989) claim, but the analytical strategy might be; homuncular functionalism may be better replaced by microfunctionalism.

Clark seems to regard Lycan's (1981a) position as still too chauvinist. Instead, Clark opts for a microfunctional account that retains a larger degree of abstraction from physical implementation and "contentful, purposive characterizations" (1989, p.35), i.e., he apparently rejects versions of functionalism that are highly teleological (without denying the enormous influence of natural selection). However, if the foregoing discussion of *S-F* theory and its status in science is mostly correct, then to reject Lycan's thesis of teleological homuncular functionalism as readily as Clark appears to is to throw one of at least two babies out with the bath water.

4.2 *Microfunctionalism and psychological explanation.*

According to Clark (1989), the success of formal descriptions of mind depend on "where you locate the grain of the input, internal state transitions, and output" (p.35). He pushes for some intermediary level between the very fine-grained, purely abstract mathematical (e.g., Turing machine table states) and the grossly detailed levels of a "semantically transparent system" (pp. 2, 18-19, 35). Neither of these two levels will provide a suitably abstracted, formal account of the necessary and sufficient conditions for ascriptions of mental

---

[7] Nevertheless, Clark (1989, p.35) does acknowledge the influence of Lycan's (1981a) defence of functionalism against Block (1978) on his own account of microfunctionalism.

states to systems, according to Clark; microfunctionalism provides our best chance of succeeding in this task. It is only through some microfunctional account that our mental state ascriptions will avoid being overly chauvinistic or liberal.

In expressing worries about necessary and sufficient conditions for the ascriptions of mental states, Clark (1989) is primarily concerned with the problems of "the project of instantiation" (p.178). As he points out, there are some very good reasons to suggest the crucial importance of distinguishing the project of instantiation from psychological explanation (Clark, 1989, pp.180-182). Most obviously, the project of instantiation is concerned with delimiting and defining the general "class of mechanisms capable of providing the causal substructure to ground rich and varied behavior of the kind warranting the ascription of mental states" (1989, p.180); whereas psychological explanation must begin by restricting its specification of this substructure to that of the brain.

Traditionally, the functionalist view has been that of the project of instantiation, with its concomitant notion of the multiple realizability of abstract functional states. On this view, the abstract functional states are the central focus of study, and the human brain is seen as one possible instantiator of these states. On the other hand, when one's approach is that of psychological explanation in and for itself, the central focus of study is some formal description of the operation of the mind-brain itself. This distinction is where I see the most fundamental difference between microfunctionalism (connectionism) and traditional computational functionalism to lie. Connectionism starts by looking at the operation of the working units of the brain (neurons), and extracts a number of operating principles and concepts

from which to build simulations and working models of the human cognitive system. Traditional computational functionalism has worked the other way round: first it specifies the computational principles, then it considers how the human cognitive system may be organized so as to operate according to these principles.

I have no doubt that Clark is right when he says that the failure to attend to this distinction between the two projects has caused a great deal of confusion in cognitive science. Indeed, I suspect that many writers on the nature of functionalism and its problems have made just this confusion. Even if connectionism, as it stands today, has serious flaws, and even if connectionist networks are not to be understood as cognitive theories (as e.g., McCloskey, 1991, argues), the microfunctionalist approach may well be pointing us in a more fruitful direction: if we want to obtain formal accounts of mind, then we may do best to abstract them from the one mind-brain that we know exists.

5. *The status of neo-functionalism.*

Despite Clark's (1989) admirable project, it is still far from certain that a system's *mere* satisfaction of some appropriate formal description will warrant ascriptions of mental states to that system (as Clark acknowledges). A microfunctional description based on some connectionist approach will likely be only one of "a set of conditions *jointly sufficient* for instantiating mental states" (Clark, 1989, p.178, original emphasis). However, even this may not be the whole story, for it is unlikely, as Clark again points out, that internal structure alone, as specified by these formal conditions, is sufficient for a system to instantiate mental states. It is entirely plausible, and I think highly likely, that

environmental, social and other semantically or intentionally characterized issues will play some part in determining the requisite conditions for mental states. Psychological explanation must include accounts of both function and structure, including any relevant factors of the physical implementation (as specified by connectionist theory, for example), relevant to the human case. Now the human case is inextricably embedded in a social context, and for many complex reasons this will certainly have some profound effects on the dispositions or capacities of humans and their patterns or types of behaviour (together, pragmatically picked out by our EMTs — also an inextricable part of the social context), and therefore on their internal brain states. Moreover, it is the operations of specific functions implemented in particular structures that allow the great variety of social contexts and activity to arise.

The social contexts of the human case may be important for full psychological explanation. But at what point does *S-F* theory depart from the specification of brain states and start dealing with interpersonal, *social* structures and functions? Although this is not an easy question (indeed, I believe the boundaries implicit in the question will remain largely 'fuzzy'), chapter 5 goes on to discuss in more detail how we might best conceive the interaction of social and brain structures and functions.

Some concluding remarks are now in order. Functional *analysis* takes us far in our quest for good (and preferably the best) psychological explanations, but not far enough. Lest we forget, functional analysis, at least as Cummins (1975, 1983) and others characterize it, is not the end of the line for full psychological explanation; a claim which Cummins (1983) himself makes, and indeed expounds on: compositional or componential analysis must also feature

in psychological explanation. However, I think Cummins, along with many others taken by the general functionalist approach, fails to give enough emphasis to the flip-side of analysis — *synthesis* — in our quest for psychological explanation. Both 'bottom-up' and 'top-down' approaches qualify as explanatory strategies, for, as I have already discussed, when our project is *explanation* of phenomena of the natural world, the structure of the implementing machinery (mechanisms of the brain, in the case of psychological explanation) must impose some limits on the postulated functions; "meringues can't carve joints" (Wilkes, 1981, p.147). The top-down strategies have been assiduously advanced in favour and ignorance of the bottom-up strategies. What these people have failed to recognize is that both types of strategy can and should exist in mutual advancement and containment (see e.g., P. S. Churchland, 1986.)

The new-and-improved versions of functionalism — homuncular functionalism and microfunctionalism — properly conceived, i.e., teleologically, within the multiple-level view of nature, are likely to be essential parts of scientific psychological explanation. Functionalism, broadly construed, is the chief and possibly only valid grand design open to cognitive psychology, and is an extremely productive and worthwhile one to boot. Nevertheless, homunctionalism and microfunctionalism are but two of the major strategies in the cooperative scientific enterprise that combines *synthesis* with analysis, *structure* with function, the *personal* with the sub-personal (etc.). Indeed, functionalism properly conceived leads inevitably to this conclusion. A conclusion that is no more applicable than for explanations of the phenomena collected together under folk psychology's 'consciousness'.

Moreover, functional analysis is an important *part* of the explanatory quest. Functionalist theories are the essence of "stage 1" of the instantiation project (Clark, 1989) and hence will help us delimit and describe, at some abstract, formal level(s), the class of mechanisms required in good psychological explanation. This is only part of the story, however. As Cummins, Lycan, and Wilkes remind us, functions and structures co-exist and are intimately related. Functions are implemented by physical stuff with certain structure, and the nature of each constrains the others. Consequently, functionalism may be better known as structural-functional theory (Wilkes, 1981).

Chapter summary:

Functionalism is the doctrine that seeks formal specifications of the causal interactions between the inputs, internal states, and outputs of a system; specifications that are abstracted as much as possible from the physical 'stuff' of implementation — Functionalism supposedly overcomes the major problems with behaviourism and the mind-brain identity theory — Turing machine functionalism is now widely discredited — Functionalist theories must not be too liberal or too chauvinist in their ascription of mental states.

Homunctionalism sets functionalism back on track by staying true to the underlying motivations: to *explain* the functioning of an intelligent system without re-introducing the very thing one wishes to explain — Homunctionalism is grounded in the analytical strategy — Homunctionalism does not flounder where earlier versions of functionalism ran aground — Homunctionalism, teleologically construed, leads one to the view that nature is multi-levelled — From this multiple-level view of nature it is evident that function and structure go hand-in-hand; consequently, functionalism may be

better known as structural-functional theory — Homunctionalism (and other good forms of *S-F* theory) can survive the sustained attacks that rely on non-standard counter-examples — *S-F* theory, like other scientific theories, is not required to explain the essence of its constructs in terms of necessary and sufficient conditions.

Microfunctionalism is another major contender for our best formal theory of the mental; one that is grounded in neurally-inspired connectionist theory — Microfunctionalism and homunctionalism need not be mutually exclusive alternatives.

Some traditional functional-computational accounts are mistaken in their attempt to put the states posited by folk psychology back into the cognitive system — Instead, *S-F* theory (homunctionalism and microfunctionalism included) should seek to explain the actual goings-on inside such systems — *S-F* theory is not a functional state identity theory; it is the "design stance" in action — *S-F* theory is thus best seen as the attempt to explain the nature and interaction of "a class of physical mechanisms capable of supporting the rich, flexible ... behavior that warrants *ascribing* mental states to the system instantiating such mechanisms." (Clark, 1989, p.178, emphasis added).

CHAPTER 3.

FUNCTIONALISM IN ACTION: COGNITIVE MODELS OF CONSCIOUSNESS.

1. *Psychology, consciousness, and functionalism.*

In the psychology of the past one hundred years or so, consciousness has gone through various stages of neglect and banishment (e.g., by the behaviourist tradition) and open embrace (e.g., by the phenomenological approaches), as well as various shades of grudging acceptance and/ or ignorance. Reporting of conscious experience — the introspective method — was the mainstay of scientific psychological investigations in the field's formative years. Nevertheless, many of these early psychologists were uneasy about the concept (Hilgard, 1977, 1980). Problems with the introspective method soon became apparent, and an alternative approach, behaviourism, became the dominant doctrine. Behaviourism essentially banished the concept of consciousness from scientific psychology. (Although it did not disappear from the scene altogether: the methods employed in psychophysics relied on an implicit acceptance of introspective report; Baddeley, unpublished; Hilgard, 1977, 1980.)

It took a considerable time after the demise of behaviourism for psychology to treat consciousness as a legitimate and respectable subject for scientific inquiry. Indeed it wasn't until the 1970s that this occurred, and the burgeoning field of cognitive science was a major player (some say *the* major player) in this resurgence of scientific interest in consciousness. Although as Hilgard (1992) remarks, "[c]ognitive psychology was not in itself a

consciousness psychology, but it opened the door for the return of an interest in conscious processes." (p.18). As a result, consciousness became a "respectable, useful and probably necessary " topic for scientific discussion and investigation (Mandler, 1975).

Underlying the increased interest of cognitive or information processing psychologists in consciousness are two primary assumptions (Shallice, 1991):

(1) the (implicit or explicit) allegiance to the philosophical doctrine of functionalism; and (2) the mapping of consciousness "onto the operation, input to, or output from one component within the system viewed from an information-processing perspective." (p.215). It is the primary task of the present chapter to discuss the second of these assumptions. But first some more needs to be said about cognitive psychology's adherence to functionalism.

Despite the gradually increasing interest in consciousness, it did seem that much of cognitive psychology could proceed quite happily with the task of mapping the mind without the need to posit any property or capacity of consciousness (e.g., Flanagan, 1990; Marcel, 1988). To be sure, it was generally assumed that the conscious processing or experience of the experimental participants often played an important role in the phenomena under study. Yet the resultant models of the cognitive system seemed not to require the postulation of any special property of consciousness other than various levels of basic awareness and control; the sort of awareness and control that suitably programmed digital computers can routinely perform. Thus "[i]t was widely noticed in many domains that the project of mapping out the complex information flows, caches, and networks constituting the mind ... could be done without bringing consciousness into the story." Computational functionalism implies "conscious inessentialism" (Flanagan, 1990, p.309).

Conscious inessentialism is an unpalatable position for many philosophers and psychologists. Flanagan (1990) does not want to advocate such a position, and neither does Baars (1988): "Whether cognitive psychology will succeed where others have not depends in part on its success in understanding conscious experience — not just because 'it is there', but because consciousness, if it is of any scientific interest at all, must play a major *functional* role in the human nervous system." (p.xv). If one wants to retain some form of functionalism and not be a conscious inessentialist, then the only conceivable option appears to be some form of *teleological* functionalism (a position that Flanagan defends). The properties, states, or events shown by science to underlie the folk-psychological notion of consciousness are likely to be *functional;* i.e., many, but not likely all, will be shown to have (or to have had) some biological *purpose* or *adaptiveness.* Consciousness, on this view, is not likely to be inessential, nor an epiphenomenon.

As it is used here, an 'epiphenomenon' is a property or entity that has no functional or design role — it may have many effects on the world, but it has no functional or causal role in the operation of the system of which it is part. So, consciousness (awareness) is an epiphenomenon, in this sense, if it exists in the physical world, and is, or is caused by, e.g., certain information processes, but does not enter into or causally influence subsequent processing. Conscious epiphenomenalism, on this view, is therefore a stronger claim than that of conscious inessentialism (Flanagan, 1990): not only is consciousness not necessary (inessential) for intelligent and purposeful behaviour characteristic of humans, it simply does not enter into it.

This version of epiphenomenalism should be clearly distinguished from the more traditional philosophical interpretation, which is that "mental

phenomena are not a part of the physical phenomena in the brain that ultimately determine our actions and behavior, but rather ride 'above the fray' " (P. M. Churchland, 1988, p.10). Mental phenomena, on this view, are caused by physical events in the brain, but have no causal effects on the world whatsoever. Computational functionalists do not (on the whole) claim to advocate this stronger, dualist position. (For more on the important distinctions between the two interpretations of epiphenomenalism, see Block, 1991; Dennett, 1978c, 1991a. Velmans', 1991, treatise on cognitive psychological studies of focal-attentive processing is the latest incarnation of such inessentialist and epiphenomenalist worries about consciousness; see chapter 4, below.)

I will have more to say about the troubles posed by conscious inessentialism and epiphenomenalism in the next chapter. The task of the remainder of the present chapter is to discuss some typical cognitive psychological (computational functionalist) models of consciousness.

2. *Consciousness in cognitive psychology: the 'consciousness module' and working memory.*

2.1 *Introduction.*

Cognitive theories that have explicitly attempted to account for consciousness in their models can be divided into four broad positions (Shallice, 1991): (i) consciousness corresponds to the operation of a single, limited capacity system, typically with control and monitoring functions; (ii) the contents of consciousness are identified with the contents of a short-term memory store; (iii) the differences between nonconscious processing and conscious processing can be accounted for by their different modes of

operation: the former is generally a parallel, primary-process, whereas the latter is generally a serial, secondary-process; (iv) the contents of consciousness correspond to the 'selector input' determining which of the parallel-acting 'action-systems' is currently most active or dominant. (It is important to note that these four positions are not all mutually exclusive. Indeed, some cognitive theories of consciousness incorporate elements from two or more of these positions.)

2.2 *The 'consciousness module' as central executive and internal monitor.*

The topic of consciousness was not dealt with directly by cognitive psychologists from the word go, but rather rode in on the back of investigations of selective attention and short-term or working memory. Some psychologists have assumed that the apparently close relationship between attention and consciousness is as good an indication as any that conscious processing can be identified with 'focal-attentive processing'. This makes the explanation of conscious processing *relatively* easy, for focal-attentive processing is explicable in the terms of information processing theories. This idea is centred on a 'two-process' theory of information processing: pre-attentive, and hence preconscious processing occurs in parallel, and is automatic, involuntary, and inflexible; whereas focal-attentive, and hence conscious processing is serial, and is voluntary and flexible (Velmans, 1991).

According to this general view, information processes that are accompanied by consciousness are at the focus of attention. Or more specifically, it is generally the *results* of such processes that are at the focus of attention and which one is aware of. Of course, as Velmans (1991) reminds us, this does not imply that *all* input and other information allocated attentional

resources will 'enter consciousness'[8]. For instance, Velmans (1991) cites the dichotic listening experiments of Triesman (e.g., 1964a, 1964b) as having demonstrated that attentional resources are utilized for some sophisticated processing (e.g., input analysis), yet nothing is available to consciousness. Certainly it is unlikely that focal-attentive processing can be *identified with* consciousness (see e.g., Allport, 1988; Dennett, 1978b; Jackendoff, 1987; Velmans, 1991). Even if one is not paying attention to something, one may nevertheless be faintly or vaguely aware of it (Dennett, 1978b). Furthermore, it does not seem possible for one to attend to something without first being aware of it (Jackendoff, 1987).

Focal-attentive processing is considered to be limited in its processing capacity. This has led many advocates of the view that consciousness is closely associated with focal-attentive processing to propose the operation of a *limited capacity central processor*. This proposed internal monitor allows us to be aware of some of our perceptual, cognitive, and bodily action processes. Only information dealt with by this processor can be available for consciousness. For instance, in the discussion of their experimental report on the components of attention, Posner and Boies (1971) suggest a link between what we can attend to, and hence what we can be conscious of, and some limited, central processing capacity. Posner and Klein (1973) also consider the possibility that consciousness is related to the operation of a limited capacity mechanism. They hypothesize that as a result of its serial operation and limited capacity, the information processes are relatively slow and are susceptible to interference

[8] For reasons that will become apparent, I use scare quotes here to disassociate myself from the prevalent "Cartesian materialist" view that there is some final 'theatre' of the mind to which conscious events 'enter', to be 'presented' or 'projected' onto the 'screen', thus forming the single 'stream of consciousness' (Dennett, 1991a; Dennett & Kinsbourne, 1992).

from other, competing tasks. Posner and Warren (1972) consider conscious processes to depend on the operation of the central processor.

A more recent version of the theory that consciousness is associated with the operation of a limited capacity processing mechanism is that of Johnson-Laird (1983, 1988a, 1988b). Johnson-Laird's theory of the conscious and unconscious mind is based on a computational framework. He postulates that 'simple consciousness' (bare awareness) can be explained by way of a high-level monitor that arises from the complex parallel processing of the brain. This monitor, or operating system, is at the top of the hierarchy of processors. The processors below the level of the operating system in the hierarchy operate in parallel, and may operate in some distributed representational manner, such as that postulated by the connectionist theories. In contrast to the processors lower in the hierarchy, the internal monitor operates serially, receiving and transmitting messages to and from these lower processors. Usually these 'messages' consist of explicitly structured symbols with a propositional content (emotional signals may be an exception; refer e.g., Oatley and Johnson-Laird, 1987).

An important distinction to be made here is that between process and product: information structures are input to and output from cognitive processes, and one cannot exist without the other. What distinguishes Johnson-Laird's theory from many earlier models is that he identifies consciousness with information structures (the information being represented), not processes. "The contents of consciousness are the current values of parameters governing the high-level computations of the operating system." (Johnson-Laird, 1983, p.465; see Jackendoff, 1987).

There is a great deal of processing that goes on of which we are not aware. Indeed it would be a ridiculous situation if we could be aware of even a small

part of the immensely complicated parallel processing at any one time; our conscious minds would be so overwhelmed that we would never get anything done. To operate effectively in real time in the real world, an organism must process a great deal of information in a very short time; hence the advantage of parallel processing. Yet action in the world, especially when it is planned action, for a particular purpose, is likely to require sustained effort and allocation of cognitive resources. A serial, limited capacity system, preferably with some control and monitoring function, is one possible solution to this engineering or design problem: The operating system's serial operation and control functions are analogous to those of the CPU in a computer; and the amounts and types of information structures of which we can be aware are restricted by the limited processing capacity of the operating system. Moreover, it is a significant operative and evolutionary advantage to have a limited capacity processor monitoring the activity of the complex parallel processing. For example, it offers a means of dealing with any information processing 'dead-locks' or 'deadly embraces' (Johnson-Laird, 1983). But a single limited capacity monitoring mechanism is not the only possible way of avoiding such computational breakdowns. It is unlikely that inside our heads is a single, limited capacity processor with control and monitoring functions. Nevertheless, it will be instructive to continue with the outline of Johnson-Laird's theory, for it is the most acceptable of all 'executive theories'.

In addition to its monitoring and control functions, Johnson-Laird's operating system constructs models of the external and 'internal' worlds. (The analogy here is with computer models; for example: models of weather systems.) Self-awareness depends on a particular mode of processing, one that has access to a partial model of itself. Self-reflection also requires access to

mental models of itself, i.e., of its own operations; and these self-reflective models can be embedded within each other. That is, self-reflective model construction is recursive — they can be constructed from the outputs of a previous model. (The significance for the present thesis of this notion of cognitive models of the self will become evident in chapter 5.)

Johnson-Laird concludes that a cognitive architecture with this type of operating system — with its separate mode of processing and recursive model construction — can account for all that we mean by consciousness (including the subjective experience of awareness, self-awareness and self-reflection).

The idea of a central processor with monitoring and control functions has not been limited to information processing theories. For instance, in Armstrong's (1968) materialist theory, consciousness is hypothesized to be "a process in which one part of the brain scans another part of the brain." (p.94). What distinguishes conscious mental states is that they are 'directed' inwards, to other mental states. Armstrong suggests that "consciousness is no more than *awareness* (perception) of inner mental states by the person whose states they are" (p.94).

Humphrey (1986) likens the internal monitor or metaphorical "inner eye" to other sense organs; but it is one that gives a view of the operations of the brain itself, rather than of the external world. It provides the organism with a useful, "user-friendly description" (p.70) of its own mental/ brain states.

Both Armstrong and Humphrey's theories presume some form of modularity of brain function. Gazzaniga (1988), a proponent of the internal monitor thesis, outlines a theory of brain function that suggests a modular organization. Gazzaniga states that the functional modules are grounded in the physiology of the brain, although it is presently beyond the capabilities of the

brain sciences to specify exactly how. The modules mediate the inputs from internal and external events received by the sensory and other input systems. The modules operate in parallel, outside of conscious awareness. Their outputs are 'considered' by an 'interpreter', which 'constructs hypotheses' as to the possible reasons for the various module responses. This hypothesis construction is centred on fitting the outputs into the overall current and "ongoing mental schema (belief system)" (p.219).

In sum, the internal monitor thesis proposes that all the major facets of consciousness can be explained by the operation of a limited capacity serial processor, sitting at the top of a highly interconnected hierarchy of parallel processors. Its primary functions are: monitoring the results of some of the parallel processing (essentially, receiving output messages from certain processors, and acting upon this information), and controlling or directing some of this complex processing (e.g., assigning certain systems of processors to do a certain job). As such, the monitor system can be likened to the general manager or central executive of the corporate cognitive system.

This view of the human cognitive system as a corporation is a common and useful interpretation of the functionalist perspective, and the internal monitor thesis of consciousness is a natural extension of the analogy. On this corporate view, a person is viewed as an organized conglomerate of interacting departments or subsystems, and, in turn, sub-departments or sub-subsystems, and so on down the hierarchy (chapter 2). Thus, as Lycan (1987) says of Dennett's characterization of homuncular functionalism:

> "I take it to suggest that we view a person as a corporate entity that corporately performs many immensely complex functions — functions of the sort usually called mental or psychological. A psychologist who adopts Fodor's and Dennett's AI-inspired methodology will describe this person by means of a flow chart, which depicts the person's immediately sub-personal agencies and their many and varied routes of access to each other that enable them to cooperate in carrying out the purposes of the containing 'institution' or organism that that person is." (p.40).

From the perspective of this analogy it seems a valid step to posit an executive or general manager in the 'penthouse suite' of the mind. This central executive sends out orders to the departments and sub-departments of the corporate mind, and receives reports on the results of their operations. On the basis of these reports and the short-term and long-term goals of the corporation, the central executive also sends out orders to the motor systems and the other departments responsible for producing behaviour, including the department for speech production (the 'public relations department').

A nice illustration of this analogy is Dennett's (1978b) model of consciousness. His first approximation of a possible high-level architecture of the cognitive system proposes a Control system or higher executive, which is responsible for a variety of control functions. For example, it allocates cognitive resources; sends directions to the *PR* system — the centre responsible for speech output; directs questions to *M* — the short-term memory buffer system; and executes a number of executive subroutines based on the answers received from *M*, e.g., 'interpreting' and 'censoring' the answer, and 'drawing inferences' from it.

Dennett's Control system is somewhat different from the central executives of alternative models, however. Dennett (1978b) regards the various facets of conscious awareness as resulting from the interaction of the control subsystems, i.e., the Control system, the buffer memory *M* and the *PR* system;

and not from the operation of the Control system alone. Moreover, "[t]he *content* of one's experience includes whatever enters (by normal routes) the buffer memory *M*." (p.169). And the two most important sources of input to *M* are from various stages of perceptual analysis (the contents of one's experience can be the results of various stages of perceptual processing), and from the *PR* system (we can experience what we say). Furthermore, the contents of *M* are output to the *PR* system or speech centre; thus we are able to report on what we experience (even if we sometimes may be at a loss to adequately express these experiences).

Dennett's emphasis on the important role of a short-term or working memory in conscious processing is an example of a prevalent view in many early cognitive theories of consciousness, and it remains at the centre of some more recent theories. It is to the discussion of working memory and consciousness that we now turn.

2.3 *Consciousness and working memory.*

The notion of a short-term or working memory, as distinguished from a more permanent or long-term memory system, has played a significant part in the history of cognitive psychology (e.g., Atkinson & Shiffrin, 1971; Baddeley, 1986). One early account of short-term memory, that of Atkinson and Shiffrin (1971), equates the proposed short-term store with consciousness; "that is, the thoughts and information of which we are currently aware can be considered part of the contents of the short-term store." (p.83). Such early theories considered short-term memory (STM) to be a unitary system: a short-term store (STS). This system was also considered to be a *working* memory: a 'work sheet' used to temporarily store the intermediate results of computations performed in

a wide range of cognitive tasks (see e.g., Baddeley, 1986, 1990).

More recent accounts of short-term memory propose instead a multi-component working memory. Baddeley (e.g., 1986, 1990), for example, submits that working memory consists of "a controlling central executive system and a number of subsidiary slave systems" (1990, p.95). Two likely candidates for the slave systems are a 'phonological loop' and a 'visuo-spatial sketchpad'.

A vast array of evidence has been accrued in support of such an analysis of working memory, from performance on tests of digit span, and serial recall of lists of numbers, words, and non-words, to experiments on mental imagery and rotation, to neuropsychological evidence from patients with memory impairments (see e.g., Baddeley, 1986, 1990). The idea of the phonological loop and visuo-spatial sketchpad, and their crucial role in determining the contents of conscious experience, is also at least intuitively plausible: visual images, those in our 'mind's eye', are by definition conscious; likewise, any material that is verbally rehearsed is conscious. So *why not* posit single, unitary systems that are responsible for the production of such phenomena? It appears that some *controlling* mechanism needs to be introduced into models of this type — something that "supervises and coordinates" (Baddeley, 1990, p.71) the slave systems — so why not posit a central executive with such control functions?

Baddeley (1986, 1990, 1992, unpublished) cites numerous experiments that supposedly offer support for a central executive; although, as he readily admits, very little is yet known about the nature of such a mechanism. These data cannot be explained solely by the operation of the slave systems, so the reasonable assumption to make is that they might be accounted for by the operation of the central executive. For example, Baddeley (1990) says of one investigation: "The authors conclude that the crucial difference between the two

groups is in working memory capacity, and since it is clearly not in the capacity of the articulatory loop, and presumably not in sketchpad capacity, the assumption is that the two groups [in the experiment] differ in the attentional capacity of the central executive." (p.139).

Baddeley (1992, unpublished) also offers some teleological or evolutionary reasons in support of the existence and functions of consciousness, and argues that these functions depend on the operation of a central executive component of working memory. For instance, sensory integration seems to be an important function that is ideally suited for consciousness, especially since conscious awareness allows for reflection upon the integrated information (allowing past experience to be used to understand present situations and to model the future). Baddeley suggests that this integration and reflection is achieved through working memory, and offers some experimental results supporting this conclusion.[9] These experimental results cannot be adequately accounted for by the articulatory loop or visuo-spatial sketchpad slave systems, so Baddeley argues that conscious awareness is one of the functions of the central executive component of working memory.

Such appeals to a central executive are, for the most part, based on the intuition that there should be *something* in there that accounts for (takes the place of) the will. Hence we get the postulation of control systems, like the 'Supervisory Attentional System' of Norman and Shallice (1980), partly because, as Baddeley (1990) says, not putting it in "leaves no place for the will, a concept that has been conspicuously missing from cognitive psychology for most of this century." (p.127). Jackendoff (1987) makes the same observation: To assign

---

[9] Although some more recent research has indicated that explanations of sensory integration may be more forthcoming from *neuroscience*. See Crick and Koch (1992) and *The New York Times*, Oct. 27, 1992, p.C10.

privilege to a part of the active mind, be it a central executive or some other high-level representation or process, "corresponds to the traditional intuition that the conscious mind is connected with the will, with the initiation and coordination of action, with the ability to make rational choices, and ultimately with one's sense of personhood." (pp.17-18).

The picture painted so far is a compelling view of mind as a neatly structured and organized system of processors going about their business with a certain amount of autonomy, but nevertheless under the watchful guidance of some 'higher' or central processor. But is it the right picture? The weight of argument and evidence tends to favour a homunctionalist-inspired modular analysis of mind. But to postulate a 'consciousness module' is pushing the corporation and computer metaphors too far, as I hope to show.

I contend that it is not necessary to posit a central executive with such important and powerful controlling and monitoring functions; it is possible that there is nothing at all like a central executive anywhere in the brain. The thesis of a central executive can be shown to be a hang-over from the Cartesian view of a locatable centre of conscious experience, a place where 'it all comes together' (e.g., Dennett, 1991a; see section 3, below). Nevertheless, some of the central executive theories are indeed detailed and significant scientific *theories*, not merely dressed-up intuition. My claim is that the intuitions motivating such accounts are misguided. We should prefer theories that do not acquiesce to such intuitions.

There is an additional task for these alternative theories: they must be able to explain the experimental data that make the executive model attractive, viz. focal-attention, conscious awareness (of e.g., sensory stimuli), the control of cognitive function (allocation of cognitive resources), and the like. These very

different accounts are taking shape in the theories of cognitive scientists like Baars (1988), Cam (1988, 1989), Dennett (1991a, Dennett and Kinsbourne, 1992), Jackendoff (1987), and Kinsbourne (1988). Nevertheless, it is not necessary for these new theories to entirely abandon the very useful and productive theoretical concepts like working memory and selective attention (see e.g., Jackendoff's, 1987, theory). Rather, if the concepts are retained, then the traditional ways of interpreting them will likely be remodelled.

3. *Possible problems with the central executive/ internal monitor view.*

The internal monitor thesis is in danger of re-introducing an intelligent, omnipotent homunculus. I don't presume that any of the internal monitor theorists would consider their version of the monitoring system to be such an undischarged homunculus, however. After all, the computer systems from which many of these models draw their inspiration carry out many complex and apparently intelligent tasks, all governed by a central processing unit that by no stretch of the imagination can be considered an undischarged homunculus.

But are the monitoring and control functions borrowed from computer technology the right sort for the brain? Is their implementation in computers ('von Neumann architecture') a suitable analogy for the implementation of control and monitoring functions in the brain? The remainder of this chapter will be devoted to exploring these questions, and to a discussion of some alternative views on more appropriate architectures of mind.

One way of avoiding a regress of intelligent homunculi is to offer an explanation of how the control system operates to achieve its proposed functions. That is, the homunculus is discharged by way of functional analysis.

The operation of the executive system is explained in terms of what it does with the information that it receives and processes: the results of computations are used in the service of further computations. This idea is central to Johnson-Laird's operating system, the Supervisory Attentional System of Norman and Shallice (1980), most working memory systems (e.g., Baddeley, 1992; unpublished), and the 'workspace' systems (e.g., Baars, 1988). The most important facet of these theories is not what the executive does, but the way that it does it.

Most modern computational-functional models of mind regard their executive system in this way, i.e., as an *information processing system.* Information structures have some use or function in the cognitive-behavioural economy; they are *functional* items. As Dennett (1991a) acknowledges, the central executive of these theories is not a "Cartesian Theater ", the place "where 'it all comes together ' " (p.107); information structures are not 'projected' on to the 'screen' of the conscious mind for us to see. Rather, the central executive receives and outputs various information structures, and in so doing performs certain processes that are deemed to be directly responsible for, or indeed *just are* the various facets of consciousness. Thus these central executive theories cannot rightly be accused of tacitly advocating "Cartesian materialism", which is the "view that there is a crucial finish line or boundary somewhere in the brain, marking a place where the order of arrival equals the order of 'presentation' in experience because *what happens there* is what you are conscious of. " (Dennett, 1991a, p.107; see also Dennett & Kinsbourne, 1992). The executive systems of computational-functionalist theories are not Cartesian Theatres, for they do not propose an end-point for cognitive and perceptual information, no final editing and projection room.

Even if these computational-functional models do not imply a Cartesian Theatre, they nevertheless retain a single, functionally identifiable processing mechanism by means of which the various facets of consciousness are made possible.

4. *The way forward: The cognitive system as multiple faculties with a global workspace.*

4.1 *Overview.*

Our original question about the utility and status of functionalism was, Is functionalism the right paradigm for scientific psychological explanations of consciousness? My answer is a resounding "yes", but only if we limit functionalist "boxology" (Dennett, 1991a) to a mapping of the cognitive systems responsible for the functions that allow us to ascribe 'first-order' mental predicates (Wilkes, 1984; see chapter 1, above). My answer is a resounding "no" if functionalist boxology is taken to the extent that some 'consciousness box' or other identifiable functional mechanism is postulated to account for the various facets of consciousness. This is the lesson learnt from the unshackling of one's theoretical approach from the vestiges of Cartesian materialism, with its concomitant notion of a Cartesian Theatre of the mind, and, relatedly, from the folk-psychological intuition that there must be some*thing* in there that corresponds to the concepts of the self and the will. We might be warranted in talking of the functions of consciousness, or at least of the functions of the properties underlying the multifarious states, events, and processes that we call conscious, but we are not warranted to speak of consciousness as arising from the operation of some identifiable, fixed, and precisely defined central executive

or module in the cognitive system. It may be useful to retain the notion of a central executive or operating system because it has heuristic and illustrative value, but this is not to assume some *actual* in-the-head control box and monitoring system.

Traditional cognitive models — those steeped in functionalist "boxology" — are being superseded by models proposing distributed collections of specialist processors, acting in parallel, all contesting for privileged computational status, with no hint of a presidential or executive processor to supervise and control this cacophony of activity. The agents (processors, modules, specialists, demons, homunculi, ...) comprising this "society of mind" (Minsky, 1985) are equipped with a working memory — a "global workspace" (Baars, 1988) — that is best viewed not in the traditional sense, as a functionally or spatially isolated system, but as a function of certain processors and the means by which they communicate.

The remainder of this chapter is devoted to a discussion of some examples of this 'new wave' of cognitive theory. It is interesting (and heartening) to note that many of these new theories have been advanced at least partly because of an uneasiness or dissatisfaction with the way traditional cognitive models have dealt with consciousness. Nevertheless, the computational and design problems these traditional models were devised in an attempt to solve still remain: How is the activity of many independent and specialized processors coordinated and controlled so as to bring about effective perception of, and behaviour in, the world? What structures, processes, or mechanisms in the computational mind could be responsible for conscious awareness, and how could they achieve this?

4.2 *Modules and Faculties: Fodor and Cam.*

A modular view of mind is the canonical form of homuncular functionalism, i.e., the explanatory decomposition of minds into relatively discrete, special purpose components or modules. As Fodor (1983) remarks, viewing the mind in this way marks a return to faculty psychology, which had its roots in the ruminations of Aristotle, and reached its apotheosis in the work of Gall in the early nineteenth century. The modern day cognitive view is considerably different from Gall's phrenology, but still adheres to its basic premise: to explain the rich and varied range of mental activity, we must postulate a range of relatively distinct cognitive and perceptual faculties. The mind-brain is a heterogenous organization of psychological mechanisms.

Unfortunately there is as yet no general consensus on the nature and function of the specialized cognitive processors (modules, faculties). Fodor (1983) considers the cognitive apparatus to be functionally subdivided into four major systems: sensory transducers, modular input systems, output systems, and a central system. The modules in Fodor's theory are the relatively peripheral input systems that receive information from the transducers, i.e., the sensory organs. These modules are defined by a number of properties, the most important being "information encapsulation", i.e., each module can use only limited types of information in its computations, and consequently there is little or no 'crosstalk' or sharing of information between the modules; and "domain specificity", i.e., each module is tied to a specific and limited set of tasks — the subject matter they deal with is restricted.

Fodor's central system ('cognition central') is responsible for all higher cognitive processes: it evaluates the outputs of the input systems in light of other relevant information such as expectations, beliefs, and goals, and is

responsible for such tasks as belief fixation, problem-solving, and thought. Unlike the input systems, and because of the functions it must perform, cognition central will not be amenable to modular decomposition (see Fodor, 1983; and e.g., Cam, 1988, 1989). Thus, on Fodor's view, the prospects for any detailed and worthwhile explanation of central cognitive processes, including consciousness, are slim. Although motivated by different initial concerns, Fodor shares the new mysterians' pessimism for the prospect of explaining consciousness (see chapter 1).

Despite Fodor's arguments to the contrary, there are some good a priori and empirical reasons for the modularity of central or higher cognition, albeit not in his strict sense of modularity (see e.g., Cam, 1988, 1989; Dennett, 1978b; Gazzaniga, 1988; Jackendoff, 1987). Although the evidence tends to favour a modular view of higher cognition, it remains equivocal. Nevertheless, I side with the modular view of higher cognition, not only because more evidence has been accumulated in support of it, but also because it is more optimistic in its outlook. If our goal is to uncover the mysteries of consciousness and other 'higher' cognitive phenomena — if we do not want to join the new mysterian camp, or their brethren — then an extended modular view of mind may be our best starting point.

More palatable views of higher cognition are illustrated by Cam's (1988, 1989) faculty theory, Jackendoff's (1987) "Intermediate Level Theory ", and Baars' (1988) "global workspace" theory. Although these theories all have their inadequacies (a detailed discussion of which is beyond the scope of this thesis), I outline their central claims so as to provide an illustration of the directions computational-functional theory is now taking: in particular, that these theories do not promote one specific processor or collection of processors as the central

executive of consciousness.

A faculty in Cam's terms "is a storage and production facility which can be characterized by its representational modality or format, the kinds of operations it typically carries out over representations in that format, and its processing connections to transducer systems, effector systems, and other faculties, in the production of behaviour." (Cam, 1989, p.167). The modules that constitute Cam's faculties are largely similar to Fodor's modules: they are informationally encapsulated and domain specific. The major difference with Cam's model is that it does not propose a strict boundary separating input analysis from cognition central; perception and cognition are not clearly segregated. (Jackendoff, 1987, takes a similar position — see pp.271-272.) Modules are more integral to the system as a whole. They are organized into clusters, or faculties, according to the perceptual or cognitive modality in which they play a part, and, therefore, according to their common representational formats, operating capacities, and roles in behaviour. These clusters are formed by selective linkages between the modules; that is, natural selection has had a hand in determining what modules are recruited for what jobs. Different faculties are responsible for different domains of experience. Moreover, "phenomenology ... can be identified with the results of operations carried out in each of these systems [faculties], and ... it is at least largely via these results that the systems achieve their cooperative efforts, so that these conscious states provide the connecting elements in the conjoint production under analysis." (Cam, 1989, p.175).

In other words, the selective linkages between faculties are formed by the conveying of information structures; i.e., the results or outputs of various faculties form the inputs of other faculties. It is this system of communication

that gives rise to consciousness (although Cam has not yet fully explained exactly how it might do so).

4.3 *Jackendoff's "Intermediate Level Theory."*

Jackendoff (1987), like Cam, proposes that the modality specificity of phenomenology is preserved ("experience is on the whole sharply differentiated by modality ", p.277), but his theory is somewhat more detailed than Cam's. Jackendoff proposes that there are multiple levels of representation within each modality, with some levels of representation being common to two or more modalities (i.e., there are 'points of intersection' along the chains of levels of representation). Interaction amongst the faculties is obtained by way of the *structure* of the information at the common levels of representation. Here Jackendoff is referring to the observation that it is the information structures that are present to awareness, not the processes that produce those structures — we cannot be aware of any actual computational activity, but only of the results of that activity. He calls this "Lashley's Observation": that *"No activity of mind is ever conscious "* (p.45).

Jackendoff's (1987) theory of the computational architecture of mind comprises: (1) two primary sorts of processors or modules: translation processors and integrative processors; (2) attentional processing that is highly intensive, and engages a "selection function" and a mechanism that directs attention; and (3) multiple levels of representation, of which an intermediate, rather than central, level of representational structure is responsible for supporting awareness.

Translation processors automatically translate information from one level of representation into information at another level. Moreover, this translation "processing is bidirectional: in perception top-down evidence refines and fills in

lower-level representations; in production bottom-up evidence guides choices in subsequent realization of the intended product ([e.g.,] speech or imagery)" (p.258). Integrative processors are required if this bidirectionality of translation is to take place, for the representations at each level within a faculty must be maintained in registration with each other. This registration requires a single, coherent structure to be produced, at each level of representation, from the information received from the translation processors at those levels.

Each faculty contains a selection function — an intrinsic function of the short-term memory for each faculty that restricts the number of structures to be considered for the privileged level of representation: awareness. The limited set of representations present in awareness at any one moment comprises those chosen by the selection function as being the most coherent.

This notion of a selection function differs considerably from the more traditional notions of selective attention (focal-attentive processing) and executive control functions of short-term memory. It is also a preferable view of the relationship between attention and awareness, given the problems with these traditional views (sections 2.1 and 2.2). Most obviously, the selection of the single most coherent or salient structure at any one moment is accomplished within or between the processing mechanisms themselves, rather than by an additional executive processor. So, for example, "perception does not send a multitude of half-baked analyses [of ambiguous stimuli] on to a higher capacity for adjudication" (Jackendoff, 1987, p.279). Rather, "[i]t is the selection function ... that is responsible for the fact that only one interpretation of an ambiguous field presents itself to consciousness at a time." (Jackendoff, 1987, p.259).

Furthermore, the other component of attentional processing, the direction of attention, is, in Jackendoff's theory, a computational process performed

within the systems of modules themselves, rather than by some higher control mechanism. Although Jackendoff has little to say about the details of the direction of attention, he sees its function as that of choosing which portions of the set of representations selected by the selection function are to undergo further intensive processing.

The attentional processes of Jackendoff's theory — the selection and direction functions — are particularly detailed and concentrated forms of processing: attentional processing will produce "more highly articulated representations, which in turn project into richer awareness" (Jackendoff, 1987, p.283). Hence the limited capacity of attention is not the result of a limited capacity central processing mechanism, but rather is the result of the computationally and physically expensive nature of this attentional processing.

Jackendoff's (1987) theory culminates in the claim that "[t]he distinctions of form present in each modality of awareness are caused by/ supported by/ projected from a structure of *intermediate* level for that modality ... " (p.298, emphasis added). He claims that most cognitive theories assume that the levels of representational structure available to the central level system (STM/ operating system/ etc.) do not directly support, or are not responsible for, consciousness. Rather, these theories, either tacitly or explicitly, propose some further level of representational structure — usually conceptual structure — that directly supports awareness. The levels of structure received by the central level system are presumably translated into this qualitatively different form of representation. For example, in Dennett's (1978b) model, the contents of the buffer memory *M*, i.e., the structures that are proposed to support awareness, appear to be of a conceptual structure (Jackendoff, 1987, p.287). This translation of faculty-produced representational structures appears to be what Cam (1989)

is trying to avoid when he claims that wholesale translation of the results output from the faculties is not required.

Jackendoff also proposes that such translation is not required: there are already some representational structures capable of supporting awareness — those at an intermediate level, somewhere between 'higher level' conceptual structure and 'lower level' peripheral (e.g., lower level perceptual) structure. For example, he argues that the levels of representational structure most closely corresponding to linguistic, musical, and visual awareness, are, respectively, phonological structure, the musical surface, and the $2^1/_2$ D sketch (see Marr, 1982). Just as we are not aware of the mechanics of computational processing, we cannot be aware of the syntactic structure of language (or of the basic low-level structures of music, vision, etc.). Although we can be aware of its conceptual structure, conceptual structure alone cannot support linguistic awareness, for "[t]his would leave no way of accounting for the fact that linguistic awareness is so sharply distinguished from visual awareness, since conceptual structure is common to the two faculties." (Jackendoff, 1987, p.290).

Thus we see how Jackendoff accounts for the modality specificity of awareness. Compare this with the likes of Cam's proposal: we can see that they have both hit on a similar idea — the modality specificity of awareness (that you are having something *described* to you, rather than actually *seeing* it, for example) is maintained by the outputs of the relevant faculties of that modality, and not by some further representational structures. These higher levels of representational structure may have a variety of important roles to play (e.g., in long-term memory), but they do not directly support awareness. Although conceptual structure may often be present in awareness (we can be aware of the meaning of an utterance, for example), Jackendoff argues that it is neither

necessary nor sufficient for awareness. Conceptual structure is not necessary for awareness, as illustrated by our ability to be aware of nonsense syllables as having phonological form, but not meaning; and it is not sufficient, as illustrated by the tip-of-the-tongue situation, where we are aware of *having* some conceptual structure about a person, place, event, etc., yet we are not directly aware of just what this conceptual structure is (we know what we want to say, yet we are missing the phonological structure — we cannot articulate the correct names and concepts). Although, as Billman and Peterson (1989) remark, this claim about the phenomenology of the tip-of-the-tongue phenomenon is contrary to one's intuition: their impression is that one *can* be aware of the conceptual structure in these cases.

I have spent a good deal of space establishing how theories proposing societies of specialist processors might account for the modality specific nature of consciousness, but what about the apparent unified nature of much experience (as when, for example, I experience this keyboard as the same object when I am both touching and looking at it)? Jackendoff's suggestion is that the unity of awareness is brought about by the central or conceptual structures — when, for example, the haptic and visual structures are "in registration with the *same* 3D model and conceptual structure, then the two modalities will be understood and experienced as simultaneous manifestations of the same object." (Jackendoff, 1987, pp.300-301).

Jackendoff's theory of the computational mind, and how it could be linked to the phenomenal mind, is very detailed and informative. Whether it is on the right track in terms of its major emphases and claims is of course a matter for empirical investigation. However, as Billman and Peterson (1989) argue, it is

likely that Jackendoff has sold himself short by arguing for a solely structural analysis of cognition — structural analysis should proceed hand-in-hand with a process analysis. Nevertheless, for the purposes of the present thesis, Jackendoff's model is a clear example of the new developments in theorizing about the cognitive underpinnings of consciousness.

4.4 *The gathering consensus: working memory as a global workspace.*

Both Jackendoff's and Cam's cognitive theories of consciousness may best be viewed in the light of a new metaphor for the mechanism responsible for awareness: the *"global workspace"* (Baars, 1988). The global workspace is a working memory, or information exchange, "analogous to a blackboard in a classroom, or to a television broadcasting station" (p.74). It provides the means for communication between the modules or faculties.

According to Baars' (1988) model, the cognitive system consists of three primary components: the global workspace, systems of specialized nonconscious processors, and "contexts" (stable, unified groups of specialized processors that have developed a privileged access to the global workspace). The specialized processors are able to broadcast messages to the global workspace, making the information available to the system as a whole.

This feature of Baars' model appears to contradict Cam's claim that communication between the faculties is limited. However, if Cam is right in claiming this, then the proposal that the global workspace makes information available to the system as a whole may still be correct — the information may be *available* to many faculties, but only a few of them may actually make use of it.

In making the information available to the system as a whole, many specialized processors previously unconnected to each other are able to cooperate to carry out a task that could not be completed otherwise. This broadcasting system is especially useful in ambiguous situations and circumstances involving novel or degraded input — circumstances where no one 'pre-wired' system of processors capable of bringing about a solution exists.

The workspace is global in two senses: the messages that are output to the workspace are broadcast globally; and the workspace is distributed globally, i.e., there is no one central broadcasting station, monitoring system, or central executive. Yet the central tenet of homuncular functionalism lives on: the cognitive system is composed of nested systems of processors. "The global workspace is the publicity organ of the nervous system; its contents, which correspond roughly to conscious experience, are distributed widely throughout the system. This makes sense if we think of the brain as a vast collection of specialized automatic processors, some nested and organized within other processors." (Baars, 1988, p.xx). Thus the global workspace theory turns the traditional central executive and monitoring theories on their heads, without giving up the spirit and principles of analytical and computational functionalism. There is no central executive with control and monitoring functions. Instead, any control and monitoring functions are carried out by the nonconscious processors themselves. Information broadcast to the global workspace can be monitored, edited and changed by these processors. Contrary to most traditional cognitive theories, control of thought and action is achieved not by an omniscient executive, but by various nonconscious processors and their access relations to the global workspace. "While the global broadcasting system is not an executive mechanism, it can be *used by* goal systems in an attempt to control thought and action." (Baars, 1988, p.353).

5. *Conclusions and future directions.*

Conventional cognitive models of consciousness have been contrasted with a 'new wave' of cognitive theories that do not acquiesce to the folk-psychological intuition of an inner centre of the will or of the self. Instead of the mind as a highly organized hierarchy of acutely specialized processors, with a boss system managing the show, the picture that is now emerging is one more akin to a "Pandemonium-style architecture" (Dennett, 1991a) — "swift generations of 'wasteful' parallel processing, with hordes of anonymous demons" (p.238) contesting for roles in 'higher order' cognitive functioning (e.g., speech production). Although the activity characteristic of Pandemonium architectures is mostly "undesigned and opportunistic" (p.241), patterns of organization and control do emerge. (Indeed, this *has* to be the case, at least for the brain, otherwise our percepts, thoughts, and behaviours would likely be as chaotic as the activity that produces them!) Just what these patterns of organization and control are, and how they are achieved, are tough but empirically determinable questions.

As Minsky (1985) has suggested, it may appear that cognitive functioning is organized and controlled by a central processor, and that there is a unity of consciousness and behaviour, but this may all be an illusion. In Minsky's theory, myriads of networks and subroutines, or "agents", compete and cooperate to achieve various cognitive tasks, with the dominant subroutine at any one time serving to unify behaviour. Other candidates for computational-functional architectures that do not posit a central executive include 'production systems' (see e.g., Anderson, 1983; Newell, 1973; and the discussion in Dennett, 1991a, chapter 9), and connectionist networks. Of course all such computational-functional systems must eventually be presumed to be

compatible with (implementable in) the extensively interconnected system of neurons that is the brain. We cannot, therefore, completely ignore the organizational and processing structure of the brain, and it will be instructive to look to see what neuroscientists have to say about how it is organized and how it might control the highly complex information processing it accomplishes. The brain does seem to contain some control structures — for example, it is widely recognized that the reticular formation and the thalamus play important roles in the control or mediation of the sleep-wake cycle and the arousal of the relevant perceptual and action systems in response to novel or emergency situations (Dennett, 1991a, p.274; see e.g., P. S. Churchland, 1988). Yet these brain structures show no hint of being an omnipotent homunculus: the control functions are fractured or widely distributed, and together they do not form a unitary boss system.

Further support for this new perspective on the architecture of the mind-brain has come from work in cognitive neuroscience and evolutionary biology (see e.g., Dennett, 1991a, chap. 7; Kinsbourne, 1988). For example, Dennett (1991a) considers that a certain amount of plasticity of brain organization must be allowed for — it is likely that the brain is able to reorganize "itself adaptively in response to the particular novelties encountered in the organism's environment" (p.184). Moreover, it is unlikely that the brain is *entirely* constructed of fixed or hard-wired, distinct modules. As Kinsbourne (1988) points out, there are no obvious "circumscribed (anatomically 'encapsulated') modules in cortex" (p.239); rather, if there are at least functionally circumscribed modules, then many of them are likely to have multiple functional roles — "multiple, superimposed functionality" (Dennett, 1991a). Hence it might be the case that various modules, at different neurological and

functional levels, are recruited sometimes as specialists and sometimes as generalists (Dennett, 1991a, chap. 7). Of course this will make the task of mapping the computational-functional architecture of the brain incredibly difficult and complex an undertaking — much more so than the comparatively simplistic conventional models of cognition would indicate. Certainly it would be extremely difficult to precisely characterize the 'contents' of conscious experience if they are taken to correspond to the information flow between certain collections of control subsystems, rather than the contents of a single processing mechanism (Shallice, 1988). Cognitive psychology needs all the help it can get — from neuroscience, computer science, and evolutionary biology, to name just a few.

While this new wave of cognitive theory shows some promise, some have expressed their doubts: Donald (1991), for instance, is worried that theories like Minsky's (1985) offer no *explanation* of how there still *appears*, from the first-person perspective, to be a central homunculus, a centre of consciousness, the will, and the self. "The homunculus is synonymous with the reflective, conscious mind, and somehow, somewhere in the protean parenchyma of mind, it must reside. *It cannot be explained away as an epiphenomenon, 'reduced' to algorithms or neuronal nets, or simply denied existence.*" (Donald, 1991, p.365, original emphasis). Despite Donald's apprehensions, however, plausible explanations of the first-person homunculus are on the horizon. Dennett (1991a), for instance, suggests that what often appears to be the functioning of a central executive system (the CPU of the von Neumann architecture of computers) is really the result of a 'virtual machine' implemented on the parallel processing of the brain. "The seriality of this machine (its 'von Neumannesque' character) is not a 'hard-wired' design feature, but rather the

upshot of a succession of coalitions of [the brain's processing] specialists." (p.254). Thus the computer analogy is still proving useful in theories of cognition, but it is becoming apparent that many theorists have been looking at the wrong features of this analogy.

I will explore this aspect of Dennett's theory in chapter 5, along with some other future directions for theories of consciousness. But first we must try to clear a path through the thickets of a sticky 'in principle' problem still confronting scientific explanations of consciousness. Central to this discussion will be the nature and status of the first-person perspective that Donald (1991) appeals to in the above quoted passage.

CHAPTER 4.

PROBLEMS AND PROSPECTS: QUALIA AND EPIPHENOMENALISM.

1. *Introduction: phenomenal consciousness and causality.*

As was mentioned in chapter 1, one of the two major characteristics of conscious mental states are their phenomenal qualities. The point of discussing the phenomenal characteristics of consciousness here is to determine, as far as possible, whether such properties pose insurmountable problems for physicalist-functionalist science (*S-F* theory).

The objections discussed below are aimed primarily at physicalist functionalism. What of objections aimed specifically at functionalist explanations of consciousness? The two classic cases are the 'inverted qualia' and 'absent qualia' objections, for they trade on the apparent weakness of that doctrine's central explanatory posits — relational properties — to explain consciousness, viz. the supposed intrinsic properties of conscious experience (Block's, 1991, consciousness$_p$[10]).

The typical absent qualia objection to functionalism runs like this: It follows from functionalism that although systems S1 and S2 behave in similar ways to, for example, harmful stimuli (they both writhe in agony when

---

[10] I use 'consciousness$_p$' as a generic term, to refer to the gamut of (whatever it is we mean by) 'awareness' or 'phenomenal experience'. Therefore the notion of 'qualia' may best be construed as a possible interpretation of consciousness$_p$; the interpretation that says: to experience X is to directly apprehend some intrinsic, ineffable quality (-ies) of that experience of X, from one's own subjective point of view (see discussion of qualia, below). As such, I may be using 'consciousness$_p$' somewhat more loosely than does Block (1991). Other interpretations of consciousness$_p$ may be possible – some will become evident in later sections.

subjected to torture, both say they are in pain, etc.), S2 may nevertheless not actually feel (experience) pain. All the relevant subsystems in S1 and S2 are functionally identical, and yet the qualia of pain is absent for S2 (i.e., $\text{consciousness}_{\text{p}}$ is inessential).

The typical inverted qualia objection to functionalism runs like this: Suppose that person P is born with a neurological defect such that when a normal person sees a red fire engine, say, P sees a green fire engine; i.e., all P's colour sensations are inverted relative to everyone else's. Still, P's behaviour would be indistinguishable from a normal person's, for P would have grown up learning all the colour words and appropriate behaviours others had learnt; for example, P would say (and believe) that the fire engine was red, even though his colour qualia would be what everybody else would consider to be green. It follows from this that functionalism could not distinguish between P and any normal person (all P's cognitive-behavioural systems are functionally identical to somebody with a non-inverted spectrum), and therefore functionalism leaves something out — to wit, the phenomenal qualities of experience. (For more detailed discussions of these arguments, see e.g., Block, 1980a; Davis, 1982; Dennett, 1988, 1991a; Kitcher, 1979; Lycan, 1987; Shoemaker, 1975, 1982, 1991.)

Whether these objections seriously impede physicalism as a whole is a matter of some debate, but this issue need not concern us here, for these arguments were aimed specifically at certain versions of computational functionalism, and I have already affirmed: (1) some good reasons to be sceptical about *those* versions of the doctrine, independently of the absent and inverted qualia objections, and (2) my subsequent allegiance to teleological structural-functional theory. (See chapter 2.) Additionally, in what follows, the primary method that these anti-physicalist, anti-functionalist arguments

employ — thought experiments — will be called into question.

Functionalism implies causality: for an entity to be functional it must have some causal status — not just having some effect in the world, but having some effect on the operations of the system of which it is a functional part. Now it is certainly the case that the typical functionalist equivalents of consciousness — be they the monitoring and control functions of a central executive, or of a variety of specialized processors — when suitably implemented, will have causal status. But what of phenomenal experience? At least four positions on the nature and causal status of consciousness$_P$ can be discerned:

(a) Phenomenal experience exists and it has a functional (causal) role in the cognitive-behavioural economy, *and* it can be captured (at least in principle) by physicalist-functionalist theory (e.g., Baars, 1988; Davis, 1982; Flanagan, 1990; Marcel, 1988; Shoemaker, 1975, 1982, 1991; Van Gulick, 1985, 1990, 1991).

(b) Phenomenal experience exists in *some* systems (viz. sentient humans), in which it will have a functional (causal) role in the cognitive-behavioural economy. *But* phenomenal experience cannot be entirely captured by functionalist theory, because functionalism is concerned *only* with causal roles; that is, there is more to qualia than their causal roles, and therefore functionalism is inadequate. Consequently, functionalism allows the possibility of absent qualia, and it may not stand up to the inverted qualia objection (e.g., Block, 1980a); functionalism leaves something out.

(c) Physicalist-functionalist theory can explain the existence of phenomenal experience; but from this third-person perspective phenomenal experience is an epiphenomenon, i.e., it has no causal role in the cognitive-behavioural economy. *However*, from the first-person perspective phenomenal experience is not epiphenomenal, and therefore the first-person and third-

person perspectives need to be given equal ontological footing (Velmans, 1991, 1992).

(d) Phenomenal experience exists but it *cannot* be explained or captured by physicalist-functionalist theory, *and* it is an epiphenomenon (the new mysterian's stance: e.g., Jackson, 1982, 1986; McGinn, 1991; Nagel, 1974/ 1979, 1986).

(e) There is no such thing as phenomenal experience over and above that captured in physicalist-functionalist theory.

In what follows, I contend that only options (a) and (e) are tenable theories of mind. By the end of chapter 5 I hope it will have become clear that I favour option (e), when interpreted in a certain way, as the most promising path for our theories of mind to take. Option (e) is open to a number of possible interpretations, from the outright rejection of phenomenal experience (of whatever sort), to the substantial claim that physicalist-functionalist theory can explain what conscious experience consists in, while nevertheless denying that there is anything at all corresponding to the conventional notion of 'phenomenal experience' (consciousness$_p$/ qualia/ 'what it is like to be X'). It is the latter rendering of option (e) that will be supported here.

A problem for explanations of the causal role of phenomenal experience (whatever it is we mean by the term), i.e., for option (a), and, in some cases (e), is: just how does this causal power manifest itself? On the one hand there is the sub-personal scientific account: All facets of consciousness are presumed to be explicable entirely in terms of the cognitive roles of brain states and events (we have rejected dualism). Therefore the causal potency of all brain states and events, and the causal origins of outward behaviour, will be explicable in terms of other brain states and events: highly complex and varied systems of interconnected neurons, patterns of spreading activation between these neurons

and neuron systems, and the 'information processing' and 'information structures' — in general, the structural-functional analysis of the brain's architecture, as promoted in the preceding chapters. On the other hand, however, we have the 'view from the inside', bolstered by the shared view of folk at large (or is it the other way round?): it is *we*, the feeling, thinking, believing, acting, *persons*, that are the causal agents of our behaviour, and it is the content of our conscious thoughts, beliefs, desires, emotions, etc., that are the reasons for our actions, and for our further thoughts, beliefs, desires, emotions, and the like. The first-person and folk views are manifestly accounts of the structure or content of mental states, rather than process accounts (see e.g., Marcel, 1988).

Which of the two accounts is correct? Can we say for sure? Can the folk-psychological and first-person accounts be explained (analyzed) from the scientific viewpoint? Or will the folk-psychological and first-person accounts be explained away? These are profound and formidable questions; questions that any science of mind must do justice to. As we shall see (initially in the present chapter, and then in more detail in chapter 5), there are some theories afoot that proffer answers to most or all these questions.

Of central import to the present chapter is the apparent gulf between the folk-psychological view of whole persons, and the structural-functional scientific view of humans as conglomerates of sub-personal machinery. One major difference is that the folk-psychological and first-personal perspectives assume a unique 'point of view', whereas the structural-functional perspective is characterized by a 'view from nowhere'. Will the scientific view always be incomplete because it appears to leave out the first-personal point of view, the 'what it is like to be X'; i.e., the uniquely private, intrinsic characteristics of

conscious experience? Or is the first-person perspective (as it is characterized in subjective reports, and in general folk-psychological beliefs) partly or completely mistaken, and therefore waiting to be assigned to the scrap heap, to be replaced by a developed science? Or indeed can some compromise be obtained, some integrative or complimentary account of the first-personal and scientific views of consciousness?

2. *Physicalism and the phenomenal mind: of bats, neuroscientists, and ineffable feels.*

2.1 *Introduction.*

Even if the global workspace models of consciousness are more on the right track than the conventional models they replace (chapter 3), it appears, at least from the first-person point of view, that something has nevertheless been left out of these attempts to explain consciousness. As Dennett (1991a) comments, "these models ... are so concerned with the *work* being done in that workspace that there is no time for 'play' — no sign of the sort of *delectation* of phenomenology that seems such an important feature of human consciousness." (p.256).

There have been some rather colourful and ingenious arguments put forward to suggest that physicalism is incomplete; that it cannot account for the subjective nature of experience. Nagel's (1974/1979) "What is it like to be a bat?" and Jackson's (1982, 1986) "What Mary didn't know" are prime examples of these arguments against physicalism, and both accept the notion that qualia exist as irreducible and essential properties of consciousness. Thus Nagel and Jackson are proponents of *property* dualism, which is essentially the position

that there are special properties of mental states that are inexplicable in terms of brain states. An implication from these arguments against physicalism is that if we are to have a scientific study of consciousness then we need not, indeed *cannot*, restrict ourselves to physical science. These arguments will be explicated below.

Dennett (1988, 1991a), in agreement with the physicalist position, wants to overthrow the idea of the existence of qualia. However, Dennett (1982, 1991a) also suggests that it is possible to study human consciousness empirically, albeit in a limited sense, and only by studying the linguistic behaviour of persons. While not wanting to deny that conscious experience is real, and has certain properties, Dennett argues that these properties are nothing like the special properties commonly ascribed to qualia. The proper business of cognitive psychology is, Dennett (1982) suggests, the internal processes, and not qualia, or beliefs, desires, and the like. In what follows, I hope to show that something along the lines of Dennett's position is the more tenable stance for the type of scientific account of mind encouraged in the preceding chapters. Indeed, not taking up this stance on qualia leads one into the dark hinterlands of science, populated by many mysterious vestiges of dualism.

2.2 *Does physicalism leave something out?*

(1). Nagel's (1974/ 1979) : "What is it like to be a bat?"

A key aspect of consciousness, according to Nagel (1979), is "that there is something it is like to *be* that organism" (p.166). Subjective experience is, by definition, unique to a particular organism. Different species of organisms, particularly those that differ considerably in their perceptual apparatus and abilities, will have different subjective experiences. Nagel's thought experiment

(originally put forward by Farrell, 1950) considers the human ability to imagine the subjective experience of a bat. He selected bats because being mammals, relatively high up the phylogenetic tree, they are likely to have *some* sort of experience; and because this bat experience is likely to be sufficiently different from our own (human) experience as to make the thought experiment particularly vivid.

The primary mode of perception for a bat is by sonar — echo-location. They do not, as we humans do, have vision as their primary mode of perceiving the world in order to judge distance, shape, motion, size, etc. They do so by correlating the outgoing sound impulses with the subsequent echoes. Because of this radical difference between humans and bats, Nagel suggests that there is a fundamental difficulty for us to imagine what it is like to be a bat. We might be able to imagine what it is like for *us* to be a bat, and to have 'bat experiences', but we are nevertheless restricted by our *own point of view*, such that no feat of imagination will allow us "to know what it is like for a *bat* to be a bat." (1979, p.169).

Thus Nagel's thought experiment supposedly demonstrates the uniqueness of the first-person point of view. The main thrust of Nagel's ensuing argument is that we cannot hope to explain subjective experience with objective, physical theories, because subjective experience is necessarily connected with a unique point of view. Objective explanations are, for Nagel, those that 'externalize' knowledge; i.e., those that strive for the abandonment of any particular, partial point of view. Ultimately, an objective theory is "The view from nowhere" (Nagel, 1986). Thus contrary to the claims of physicalism, it will be impossible for an objective neuroscience to account for qualia. Further, Nagel (1986) asserts that the "subjectivity of consciousness is an irreducible

feature of reality ... and it must occupy as fundamental a place in any credible world view as matter, energy, space, time and numbers." (pp.7-8).

Nagel's argument is an attack on reductionist versions of physicalism. (Note that his conception of reductionism here is fairly broad — it includes, for example, all those attempts to explain consciousness in terms of nonconscious mechanisms, events, or properties.) Reductionism relies on an analysis of what is to be reduced, according to Nagel: every phenomenon held to be a feature of the concept that is to be analyzed (viz. the subjective nature of consciousness) must be accounted for in this analysis. If something is left out, then the reduction will be unsatisfactory or inadequate — it will not fully explain the concept in question.

As Armstrong (1981) points out, qualia (the "secondary qualities" of sensory experience) appear to be unanalyzable. As we experience them, secondary qualities — the blueness of an object, for example, or its felt texture — strike one as elementary, indivisible properties. If qualia are elementary properties, then they cannot be analyzed (explained), for this would merely reintroduce the very thing one is trying to explain (the subjective nature of experience), resulting in a circular explanation.

So what are we to make of Nagel's argument? Three significant flaws have been noted: that a failure of our imaginations cannot *prove* physicalism is inadequate; that Nagel misrepresents the nature of scientific explanation; and that the argument depends upon an equivocation on the meaning of 'knowledge'. I now turn to a review of the first two faults, and will consider the third as it applies to the next thought experiment, Jackson's (1982, 1986) "What Mary didn't know".

As has been observed by Copeland (unpublished), for example, the "What is it like to be a bat?" thought experiment is misguided in its line of attack on physicalism. All that the thought experiment proves is that our imagination will not be able to give us the true essence of 'bat experience', and that subjective experience means experience from the 'subject's' point of view. It does not prove that bats qua bats are not purely bio-physical entities, and that therefore physicalism might be leaving something out. "We should be interested in what we can know about the bat's consciousness (if any), not whether we can turn our minds temporarily into bat minds. ... There is at least a lot that we can know about what it is like to be a bat, and neither Nagel nor anyone else has given us good reason to believe there is anything interesting or theoretically important that is inaccessible to us." (Dennett, 1991a, p.442).

What Nagel's thought experiment does have is a certain amount of intuitive appeal: that the third-person view of science cannot capture just what it is like to be a bat, a person, or other suitably advanced cognitive system, because *what it is like* to be a bat, a person, etc., is unique to any one particular individual of that species (although there will surely be similarities, due to e.g., similar perceptual apparatus). To put it another way, 'one has to be there' to *know* what it is like; no mere statement of what it is like will suffice.

If physicalism is to survive this challenge, we must attempt to undermine this intuitive appeal with some sound, empirically supported, theoretical alternatives. Some such attempts have been embarked on, and these will be introduced in section 2.4. In the meantime, there is a further problem with Nagel's argument that deserves attention. This will then be followed by a consideration of a possible way of avoiding Nagel's conclusion.

Rosenthal (1986) claims that Nagel misrepresents the nature of scientific explanation. Nagel (1974/1979) holds that reductionism requires an analysis of

everything that is to be reduced. If something is left out of the analysis, then the reduction cannot go ahead — the concept in question cannot be fully explained. As stated above, all attempts to explain consciousness in terms of nonconscious states (including the structural-functionalist accounts discussed in the present thesis) qualify as reductionist analyses, in Nagel's view. Nagel's argument fails to impugn these attempts at explaining consciousness, however, for that is not how scientific explanation proceeds. "Explanation, in science and everyday context alike, must generally proceed without benefit of complete conceptual analysis." (Rosenthal, 1986, p.352).

As discussed in chapter 1, some form of analysis is often the primary mode of scientific explanation, especially psychological explanation. Relatively comprehensive conceptual analysis may eventually be achieved by the process of scientific explanation, via a variety of other forms of analysis (e.g., property analysis, compositional analysis; see Cummins, 1983), but thorough conceptual analysis does not *precede* scientific explanation. Conceptual analysis may need to be carried out to some relatively minor extent prior to scientific explanation: namely, analyzing the general concept one is investigating — consciousness, for example — so as to delimit the phenomena to be explained, and to avoid ascribing predicates applicable only to whole persons to parts of persons (brains, functional modules, etc.). (See e.g., Kenny, 1984, p.134). But this limited scope of prior conceptual analysis does not require us, as Nagel claims, to precisely circumscribe the meaning of 'what it is like to be X' without undermining our intuitive thoughts about such matters (that phenomenal experience, 'what it is like to be X', cannot be fully explained in the languages of the physical sciences). Nagel claims that "[w]ithout some idea ... of what the subjective character of experience is, we cannot know what is required of

physicalist theory." (1979, p.167). I agree, up to a point: *pace* Nagel, physicalist theory is not restricted to accounting for only those concepts that have been exhaustively analyzed. As Dennett (1991a) says, "It is premature to argue about what can and can't be accounted for by a theory until we see what the theory actually says." (p.71).

The intuitive impossibility of explaining the subjective nature of experience in the terms of physicalist science should not blind us to the prospect of it actually being possible. As was pointed out in chapter 1, we *are* permitted to draw explanatory links (inferences) between conscious experience (viz. sensory awareness) and brain properties, when we can establish they are reliably linked, because in such cases we have a prior commitment to the existence of conscious experience (the subject has reported it). That is not to say that we should attempt to *reduce* (identify) the introspectively reported contents of consciousness to properties of the brain. But, *pace* Paul M. Churchland (1985), this is not exactly what Nagel is advocating (McCulloch, 1988, p.7). Nevertheless, this characterization of Nagel's position is not entirely off-track: Nagel claims that all attempts at reductively explaining consciousness (i.e., those that attempt to explain it entirely in terms of nonconscious entities) cannot succeed; "there is no reason to suppose that a reduction which seems plausible when no attempt is made to account for consciousness can be extended to include consciousness." (Nagel, 1979, p.167). But attempts *are* made by some physicalist theories to account for consciousness, and this does not, contra Nagel, require a prior analysis of the subjective nature of experience, of 'what it is like to be X'.

In reply to Nagel's pessimistic conclusion about the success of physicalism, this is what a physicalist might say: But what if this apparent

simplicity or unanalyzability of phenomenology is epistemological, not ontological? (Armstrong, 1981). What if, in other words, the apparently unanalyzable, elementary nature of secondary qualities is *merely* apparent, merely a matter of our awareness of them, and not of the way they are? But what exactly does Armstrong mean when he says this? Presumably he means that the way things are in the world is not always evident as such to perceiving individuals; and, as Armstrong is a materialist, I assume that he would say that it is the task of science to reveal just what the true natures of things in the world are. If science is any good at its job, then it should eventually uncover the ontological nature of secondary qualities of experience. (Armstrong is not a new mysterian, by the way, as is Nagel.) But this is precisely the sort of explanation Nagel takes exception to: no third-person account will be able to explain exactly how things appear to us, exactly what it is like to be a sensing, perceiving, feeling organism.

As we have seen, Nagel does not present a very secure case against physicalism (his argument is faulty, and he misconstrues the nature of scientific explanation). As a number of Nagel's critics have tried to show, no amount of *a priori* argument of the likes of his thought experiment will establish that physicalism cannot explain the first-person nature of experience. We will just have to wait and see whether physicalism comes up with the goods. Still, the new mysterians have not given up the fight (see e.g, Jackson, 1982, 1986; McGinn, 1991).

To recapitulate: In one corner are the friends of qualia (of whom the most conspicuous are Nagel and the other new mysterians), while in the other corner sit the optimists — those attached to some version of physicalism. Nagel claims that the subjective 'feels' of experience — the way things appear *to us* — are

elementary, irreducible features of the ontology of experience. Any account that leaves out the first-person perspective (viz. physicalist science) will not fully explain consciousness. That is, there are facts about experience that no objective theory can describe. The optimistic physicalists are holding out for some future time when the brain and cognitive sciences will be near enough complete, at which point *all* facets of consciousness will be explicable. Or will they? Let us look at Jackson's (1982, 1986) thought experiment, which illustrates an argument with a conclusion similar to Nagel's.

(2). Jackson's (1982, 1986) "What Mary didn't know".

Jackson argues that the thesis of physicalism is incomplete because it cannot account for certain features of the subjective experience of sensations; there are facts about experience that physicalism cannot describe. He presents this argument in the form of a thought experiment concerning a fictional character of the future named Mary. Since birth, Mary has been confined to a totally black-and-white room. Her primary source of communication with the outside world is via a black-and-white television. Everything she has seen all her life has been devoid of colour. She receives an extraordinarily advanced education via the television, such that at the end of the course she knows all there is to know about the nature of the physical world. As Jackson (1986) says, "If physicalism is true, she knows all there is to know." (p.291).

Then one day Mary leaves the room (or is given a colour television). She is now able to *experience* what it is like to perceive colour. Jackson views this new type of experience in Mary's life as the *learning* of something new; the gaining of some *further* knowledge of the world. Prior to the introduction of the colour television, Mary knew *all there was to know* about the physical world, including

the neurophysiology of colour vision. Yet on perceiving colour Mary gains some extra knowledge.

Jackson concludes from this that physicalism is false, because it cannot account for this extra knowledge that Mary has gained. However, in the latter version of this argument (1986), he adds that the problem for physicalism is not the additional knowledge that Mary gains about her own experiences, but it is "the *knowledge about the experiences of others*." (p.292, original emphasis). Jackson's objection to physicalism is not that Mary learns something on perceiving colour for the first time, but that "she will realize how impoverished her conception of the mental life of *others* has been *all along*." (p.292). She knew all the physical facts of colour vision and other people's experiences, and so according to physicalism, that is all there is to know. If there was something to know prior to receiving the colour television then it is not a physical fact. But the claim of physicalism is that *all* things are physical. This is the contradiction that Jackson believes is an insurmountable problem for physicalism.

The major criticism levelled at Jackson's argument (see e.g., P. M. Churchland, 1985, 1988; Levin, 1986) is that it depends on an equivocation between theoretical knowledge (knowledge of the facts) and experiential knowledge or ability (to know what it is like to see red). (As it turns out, Nagel's "What is it like to be a bat?" also depends on this equivocation.) The extra knowledge that Mary gains is simply knowledge in the sense of personal acquaintance. This is true for both the knowledge that Mary gains about her own new colour experiences and for the knowledge she gains about the experiences of others. Jackson argues that the knowledge that Mary gains by her acquaintance with coloured objects is not knowledge of the physical facts about those experiences (rather, it is an 'acquaintance' with the qualia of those

experiences); and because physicalism requires this knowledge to consist of physical facts, Mary couldn't have known *all* the physical facts about colour vision prior to her seeing colours for the first time: therein lies the trouble for physicalism. However, *pace* Jackson, physicalism does not require that knowledge consists entirely in the learning of a series of physical facts. We can maintain that 'knowledge by acquaintance' exists, and is an *entirely* physical state or event, while rejecting the stipulation that to acquire this sort of knowledge is to learn one or more facts about the physical world. The physicalist is not committed to the claim that having complete factual knowledge about the brain may allow one to have experiential knowledge (the experience of seeing red, for example), as Jackson alleges. "In sum, there are pretty clearly more ways of 'having knowledge' than just having mastered a set of sentences, and the materialist can freely admit that one has 'knowledge' of one's sensations in a way that is independent of the neuroscience one may have learned." (P. M. Churchland, 1988, p.34).

Moreover, as Paul M. Churchland (1985) and Dennett (1991a) rightly comment, we cannot readily imagine, as Jackson's thought experiment asks us, just what it would mean to have *all* the physical facts about the brain. How can we be so sure that a 'complete' scientific understanding of the brain will not allow us to have experiential knowledge of the sort Jackson alludes to? We have little or no idea just what a 'completed' neuroscience or *S-F* theory would look like, let alone what we would and would not be able to come to 'know'. It is simply too early to tell. Jackson's thought experiment, like Nagel's "What is it like to be a bat?", relies on the failure of our imaginations to provoke the intuition that the subjective nature of consciousness cannot be explained by physicalist science — that what does not *seem* to be the case *is not* the case.

While the above rebuttal of Jackson's argument for the incompleteness of physicalism appears to be sound, it fails to show that knowledge by acquaintance does not involve some form of acquaintance with the appropriate qualia. Section 2.3 summarizes some attempts to quash qualia.

We have seen that some of the major arguments of the new mysterians and other friends of qualia against physicalism are ill-founded. If we are to be good physicalists, then we must suppress any remaining worries raised by these objections based on appeals to 'phenomenal qualities' or 'qualia'. To do this, we must accomplish two tasks:

(1) Eliminate qualia — that is, show that the qualities of experience do not have the putative special properties that constitute 'what it is like to be X'; that there really are no such things as qualia. (We don't need to deny that we have subjective experiences — consciousness_p — however; all we need to do is redefine their properties in entirely physical terms.) (Section 2.3.)

(2a) Show how it is possible to correlate these redefined properties of experience with facts about the brain; or, what amounts to a stronger position, (2b), show how neurophysiological facts about the brain can account for these properties. (Section 2.4.)

2.3 *Qualia: phenomenal qualities or chimera?*

What follows is a summary of some of the main points raised by those who wish to refute scepticism about an adequate science of mind based on appeals to unassailable, mysterious phenomenal qualities. Points (i) to (iii) are made in Kitcher (1979), and points (iv) to (vii) are gleaned from Kitcher (1979), Dennett (1988)[11], Levin (1986), and Wilkes (1988b):

---

[11] Dennett (especially 1988, 1991a) and Kitcher (1979) provide two of the most sustained and detailed attacks on qualia. Other notable onslaughts include P. M. Churchland (1985, 1988), and Wilkes (1981, 1984, 1988a, 1988b). Arguments

(i) Questions of the sort, 'What is it like to perceive X (e.g., see red)?' refer to *perceptual states*, not the perceived objects. (Although this distinction between qualities of experience and qualities of experienced objects is not often made; the confusion of the two interpretations leads to many of the problems with the notion of qualia; see Harman, 1989; Wilkes, 1988b.)

(ii) Questions of the sort, 'What is it like to perceive X?' can be given two readings: what it is like to perceive X is similar to what it is like to perceive A, B, C ... "in indefinite and unknown ways" (Kitcher, 1979, p.124); or, what it is like to perceive X is the result of a definite quality of the perception of X. But the former, relational or "comparative" interpretation of the question is no challenge to physicalist theories of mind; so those who question the capacity of science to account for phenomenal qualities must be referring to the latter, "positive" interpretation.

(iii) Therefore, phenomenal qualities (qualia) are definite qualities of perceptual states.

(iv) Defenders of phenomenal qualities regard these qualities as ineffable, intrinsic, private, and directly knowable; in short, mysterious. "The infallibilist line on qualia treats them as properties of one's experience one cannot in principle misdiscover, and this is a mysterious doctrine ..." (Dennett, 1988, p.55).

(v) The standard objections to scientific accounts of what it is like to perceive X assume that when we are in the state of perceiving X we can be aware of that state and therefore *directly* note its distinctive, definite qualities ("reflectionism"; Kitcher, 1979). However, there are problems with this notion

---

for and against qualia are presented in e.g., Block (1980a, 1991); McCulloch (1988); and Shoemaker (1975, 1982, 1991), who all come out in support of *some* form of qualia or phenomenal experience.

of 'directness', where 'direct knowledge' is usually taken to mean that it is 'non-inferential' and 'immediate'. Shoemaker (1975), for example, argues that *all* awareness or knowledge is the result of a (not necessarily conscious) description or inferential process. Moreover, "[b]oth the notion of 'immediacy' and that of 'without inference' become dubious when we allow ... for the existence of tacit knowledge and non-conscious, subdoxastic states[12]: not all that counts as inference need be conscious, so we may be unaware of many of the inferential steps in our thought processes." (Wilkes, 1988b, p.175). Other worries with this notion of 'directness' include Levin's (1986) complaint that 'direct knowledge or acquaintance' ("direct recognitional capacity") is ambiguous: it can refer to either the possession of a concept, or to "having the wherewithal to apply it." (p.248); and the vagueness and incoherency of the notion as illustrated by Dennett's (1988) intuition pumps (see below). Thus we should at least be wary of appeals to the 'directness' of knowledge by acquaintance; further, it seems that the 'directness' referred to can be nothing like what the friends of qualia assume it to be.

(vi) 'Directness' or 'immediacy' are not the only vague and ambiguous characteristics of qualia. There are a number of different interpretations of 'privacy', all of which seem fairly innocuous (e.g., Wilkes, 1988b, p.175), and nor do they indicate the 'specialness' or unassailability of qualia. We are not in any important sense 'privileged' about, or 'incorrigible' with respect to, our sensory experience, as revealed by, for example, cases of subliminal perception and 'blindsight' (Wilkes, 1988b, p.177). Moreover, it seems that the qualities of experience are only "*practically* ineffable" (Dennett, 1988, p.68); 'ineffability' does not signify the 'specialness' or unassailability of qualia either.

---

[12] Subdoxastic states are psychological states that causally contribute to belief states, but which are not themselves beliefs; see Stich, 1978.

(vii) If the characteristics of phenomenal qualities, whatever they are, are not as special and unassailable as the friends of qualia make out, then it is unlikely that they need remain mysterious. Thus we should prefer theories that purport to *explain* the way we detect perceptual states and their qualities over theories that merely state that we do and then refer to these qualities as mysterious. That is, we should prefer a relational or comparative interpretation over reflectionism (see point (ii) above); and therefore we should be distrustful of appeals to the 'intrinsicality' of phenomenal qualities (where 'intrinsic' is taken to mean 'non-relational'). "Why not give up intrinsicality as a second-order property altogether, at least pending resolution of the disarray of philosophical opinion about what intrinsicality might be? Until such time the insistence that qualia are the intrinsic properties of experience is an empty gesture at best; no one could claim that it provides a clear, coherent, understood prerequisite for theory." (Dennett, 1988, p.68).

If reflectionism — the pro-qualia, pro-'what it is like to perceive X' standpoint — is an unsound and inaccurate examination of conscious perception, then why is it such an attractive position? Primarily because, according to Dennett (1991a), Kitcher (1979), and Wilkes (1981, 1988b), the thought experiments upon which the qualia lobby rely dupe us into accepting the intuitive gulf between the subjective qualities of perceptual states and the objective explanations of those states and their qualities. We should thus be sceptical of the success of such thought experiments: "I suggest that we become puzzled about phenomenal qualities by engaging in thought experiments." (Kitcher, 1979, p.127). "Our ability to tell when we are in perceptual states and when we are aware of sensations, *and our capacity for imagining being in those states* are always available to mislead our theorizing about sentience." (Kitcher,

1979, p.129, emphasis added).

Nagel's "What is it like to be a bat?", and Jackson's "What Mary didn't know" thought experiments, for example, deceive the unsuspecting reader by pandering to their intuitive feel about the case in question without fully clarifying just what it is that physicalism supposedly leaves out. However, as Wilkes (1981) comments, "The onus is surely upon the objector [to *S-F* theory's capacity to deal with phenomenal properties] to try to spell out more precisely just what he feels has been left out; until he does, the objection cannot be decisive." (p.165).[13]

Dennett's (1988) attempt to show that there are no such properties as qualia relies on a series of thought experiments. Yet one could object that the criticisms Kitcher, Wilkes, and Dennett himself raise about thought experiments are equally applicable to Dennett's case against qualia. However, as Dennett (1991a) is quick to point out, his thought experiments (he calls them "intuition pumps") are not, for the most part, designed to convince the reader that something is in principle impossible, but rather to open one's eyes to the realms of the possible. Dennett seeks to expel the common conception that the so-called qualia of subjective experience are "ineffable, intrinsic, private, directly or immediately apprehensible" (p.47)[14]. By describing a series of thought experiments, he attempts to 'pump' the reader's intuitions about subjective experience towards the view that its properties or characteristics are really nothing like qualia; in effect, "there simply are no qualia at all." (Dennett, 1988, p.74). Their purpose is to rid the reader of many of her preconceived notions of the special nature of qualia. Unlike the thought experiments of the pro-qualia

---

[13] A more sustained and detailed examination of the efficacy and validity of thought experiments is given in Wilkes (1988b).

[14] Dennett (1988, pp.47-48) is quick to allay any suspicion that he might be setting up and knocking down a straw person here.

lobby, Dennett's thought experiments do not rely on the nuances and failures of our imaginations to persuade us that something is in principle impossible: that the apparent hidden, elusive nature of our minds will forever lie beyond the reaches of a physicalist science.

The "Brainstorm machine", for example, is an extension of the inverted spectrum thought experiment. The Brainstorm machine is a hypothetical device that transfers person A's visual experience into person B's brain, such that if A has a colour spectrum that is inverted relative to person B's, B, with eyes closed, will experience the colours of the things which A is looking at as opposite to what B is accustomed to. (A might be looking at a fire engine, saying it is red, when *this same experience* appears green to B.) But what if, Dennett asks, the cable connecting A's and B's brains is inverted 180 degrees such that B now experiences A's visual sensations with the 'correct' colour qualia? We would be confronted with a dilemma: we would not be able to decide which is the 'right' orientation of the connecting cable. Whatever way you conceive the 'experiment', in the end it would still require some relative normalization of the two people's reports of their subjective experience. But this simply takes the argument full circle back to the original starting point: the experimenters would not be able to tell whether A's qualia are inverted relative to B's.

Thus the Brainstorm machine intuition pump is an attempt to show that the inverted qualia thought experiment, even if we allow for extremely advanced technology, is inadequate. Nobody can come to know just what qualia result from what perceptual encounters, *including the person whose qualia are in question*, as Dennett (1988, p.50) demonstrates by presenting the "alternative neurosurgery" intuition pump — his own version of 'the *intra*personal inverted spectrum'. Dennett asserts that these thought

experiments serve to demonstrate that *if* qualia exist, then they are certainly less directly accessible than many believe. Both intersubjective and intrasubjective comparisons would not be able to tell us whether our own qualia have been inverted.

Still, intuition pumps, if they are used wisely, are only the preliminary steps in a long journey of scientific understanding. "If they help us conceive of new possibilities, which we can then confirm by more systematic methods, that is an achievement; if they lure us down the primrose path, that is a pity." (Dennett, 1991a, p.440).

So what are we left with if there really are no such properties of experience as qualia? Dennett's (1988) intuition pumps lead him to propose that the properties of conscious experience are extrinsic, relational properties. (See also Marcel, 1988.) Relational properties are, or can become, matters for public scrutiny; they are amenable to third-person analysis. Contrary to what a number of people have seemed to assume, extrinsic (public), relational properties do not need to be grounded in some more fundamental 'intrinsic' (especially 'in-the-head') properties. Despite many attempts to ascertain otherwise, the notion of intrinsic properties (of whatever sort) remains mysterious and incoherent (see point (vii) above).

2.4 *Bridging the gulf between phenomenology and physicalism.*

The new mysterians' stance, as exemplified by Nagel's and Jackson's thought experiments, is not only pessimistic about the prospects of physicalism, but unduly so. Even if there are something like qualia as they are customarily conceived — even if 'what it is like to be X' is a unique property of conscious beings — it is certainly not 'obvious', *pace* Nagel and Jackson, that no third-

person scientific account can explain all there is to know about them. Similarly, *pace* Velmans (1991; see section 3 below), we have no good reason to assert that no third-person scientific account can explain all there is to know about consciousness$_P$, whatever its nature. To be sure, deciding whether a physicalist theory *can* explain the properties of subjective experience is likely be a difficult task, especially if we are to convince the sceptics. But one striking hallmark of scientific theory is that it is able to show (and indeed, has shown) that what may *appear* to be impossible is instead possible.

Nevertheless, it is likely that "if we are to give a fair hearing to a theory, in the face of such scepticism, we will need to have a neutral way of *describing the data* — a way that does not prejudge the issue." (Dennett, 1991a, p.71). Dennett's (1982, 1991a) suggestion for such a neutral method for bridging the gap between first-person phenomenology and third-person science is "heterophenomenology": an impartial transcription and interpretation of the subject's verbal reports about what they experience, and of any other behaviours designated as suitable substitutes for, or adjuncts to, verbal reports (e.g., button-pushing in an experiment). The interpretation in this method is analogous to a reader's interpretation of a work of fiction: the heterophenomenologist interprets a 'work of fiction', of sorts — a picture of the world as described by the subject. Once this interpretation is complete — once a set of "intentional objects" in the subject's "heterophenomenological world" has been extracted ("mental images, pains, perceptual experiences, and all the rest") — "theorists can then turn to the question of what might explain the *existence* of this heterophenomenology in all its details." (Dennett, 1991a, p.98). Any scientific theory of the mind must account for these experiential states, and a person's beliefs about these states, for "[t]he heterophenomenology exists —

just as uncontroversially as novels and other fictions exist." (Dennett, 1991a, p.98).

Flanagan (1990), in criticising the new mysterians' position, alludes to the same method: Because both facts about the brain and facts about consciousness (including those facts as seen from the subject's point of view) require explanation, we must eventually "infer that the constellation of a certain set of autophenomenological reports of restricted range ('tastes sweet') correlate with certain sorts of brain activity ... and we infer, because of an overall commitment to naturalism, that the latter explain the former." (p.338).

What could these certain sorts of brain activity be? That is, how does the brain represent features of the world as picked up by our sense organs such that they 'correlate' or 'translate' to 'tastes sweet', 'looks purple', 'smells stale'? One plausible neuroscientific model, as summarized by Paul M. Churchland (1988, 1990), suggests that the brain represents features of the world in an appropriate "state space", and performs computations on these representations. Each sense modality can be partitioned into the different levels of neural stimulation that go to make up the overall neural 'fingerprint' of a particular item as represented in that sense modality. So for example, there are four levels of neural stimulation for taste, corresponding to the four distinct types of receptor cells; and there are three levels of neural stimulation for colour, corresponding to the three kinds of receptor cells in the retina. These levels of stimulation can be expressed as the dimensions or axes of a state space (the dimensions for colour, for example, would form the three axes of a cube); and the individual values of each of these levels for any one distinct item (the red-orange light of the setting sun, for example) will be expressed as magnitudes along these axes, together forming a unique "sensory coding vector" for that item.[15] (As such, state

spaces are excellent examples of abstract functional items; P. M. Churchland, 1988, 1990 provides a plausible outline of how they might be neurally instantiated.)

The patterns of neural stimulation that identify a unique coding vector in a given sense modality are essentially patterns of "spiking frequencies" across the relevant neural channels. Churchland (1990) believes that this neural vector model therefore lends support to a reductionist position on qualia, such that: "[t]he 'ineffable' pink of one's current visual sensation may be richly and precisely expressible as a '95Hz/80 Hz/ 80 Hz chord' in the relevant triune cortical system"; or that "[t]he 'unconveyable' taste sensation produced by the fabled Australian health tonic, Vegamite [sic], might be quite poignantly conveyed as a '85/80/90/15 chord' in one's four-channelled gustatory system" (p.361). However, as Flanagan (1990) makes clear, "[t]he thesis is not that for each different vector there is a distinct quale" (p.330); rather, the thesis holds that *only if* there is a difference in the magnitudes of neural vectors will there be a difference in "experiential sensitivity", not 'if and only if'. Thus "informational sensitivity" is necessary, but may not be sufficient for, experiential sensitivity.

While this account of the neural representation of the physically specifiable qualities of sensory and perceptual states seems plausible, I'm sure the friends of qualia will not be satisfied, for it does not seem to show how the patterns of spiking frequencies can give rise to the 'intrinsic phenomenological qualities' of an experiential state. As Dennett (1991a) says, talk of colours, sounds, smells, and the like, being 'coded' by the brain by way of vectors of

---

[15] Note that it may be possible to extend this neural vector model to face recognition and motor output, amongst other representational tasks; see P. M. Churchland (1988, 1990).

magnitudes can lead one to suppose (mistakenly) that there must be some subsequent 'decoding' back into the appropriate colours, sounds, and smells. But this supposition cannot be right for who would the recoded colours, sounds, and smells be for, except for a 'ghost in the machine', undischarged homunculus? Although the neural vector account of sensory and perceptual qualities remains neutral about the so-called intrinsic qualities of experiential states, this may nevertheless be to its advantage, given the doubts raised about any such 'intrinsicality' (see above). Moreover, the neural vector account has the added advantage of preserving the structural relationships between the physical magnitudes of the sensory stimulations (Dennett, 1991a, p.350).

Whether or not the sensory coding vector account is an accurate picture of how the brain represents the physical properties of stimuli relevant to the qualities of experiential states is yet to be confirmed beyond reasonable doubt. Nevertheless, it appears to be at least on the right track, and there is a growing body of neuroscientific research backing it up. Although, as Churchland and Sejnowski (1989), and Dennett (1991a) remind us, theories of how nervous systems represent the world need to be informed and validated by an evolutionary and developmental account of nervous systems, so a 'purely' neuroscientific account will not be able to give us the one true picture of how we are to 'quine'[16] qualia (or, in Paul M. Churchland's, 1985, 1988, 1990 view, how we are to 'reduce' qualia).

Moreover, even if the theory of neural representation as described by Churchland is the most promising account of how sensory properties are represented in the brain, it is not yet a principled account of neural representation, even if it is more than likely a significant step along the road to such a principled account. For one thing, it is not clear why a '85/80/90/15

[16] See Dennett (1988, p.42).

chord' *is* the taste *of Vegemite,* rather than, say, coffee, or indeed why it is even a taste at all.

Does the sensory coding vector theory show how the qualities of sensory and perceptual experience are realized in the brain? On one level — the third-person perspective of science — it shows some promise, for it may be able to explain how the putative physical features of a stimulus (those relevant to our sensory and perceptual experiences of an item) can be represented in the brain. But on another level — the first-person perspective, when infected by notions of 'intrinsic' properties — the sensory coding vector theory does *not* appear to show how the qualities of experience are identical to certain activities of the nervous system; the first-person perspective on the explanatory gap between neuroscience and introspection looms large.

3. *Epiphenomenal phenomenal experience?*

Some have argued (e.g., Velmans, 1991) that whatever exactly consciousness_P is, it looks increasingly likely, at least from the third-person perspective of science, that it plays no part in the causal scheme of the mind-brain and behaviour. This section uses a critique of Velmans' argument for this conclusion to illustrate: (1) once again, the pitfalls of grounding one's theory in unsubstantiated intuitions; (2) the frequent misrepresentation of the notion of epiphenomenalism; and (3) the often implicit Cartesian materialist underpinnings of many theories of mind.

Velmans' (1991) article is the latest detailed exposition of modern day epiphenomenalism as exemplified by conventional cognitivism. What is unique about Velmans' account is that he considers the first-person perspective — from

which epiphenomenalism is false — to have equal ontological footing as the third-person perspective — from which epiphenomenalism appears to be true. The apparent paradox that arises from this position is one "that any complete theory of the mind must offer to resolve." (Velmans, 1991, p.713). Velmans considers neither the third-person or first-person accounts as privileged.

To demonstrate the epiphenomenalism of the third-person perspective, Velmans surveys many cognitive psychological studies on focal-attentive processing (e.g., divided attention studies, such as dichotic listening tests). Various experimental findings indicate that some sophisticated processing can occur without reportable consciousness[17] (e.g., the processing of the meaning of familiar stimuli), or without conscious control (e.g., perceptual analysis of well-known stimuli, adaptive responses to the environment), or without prior consciousness (e.g., the learning of novel stimulus patterns). What these results suggest is that consciousness$_p$ is not necessary or required for any information processing; rather, awareness appears to follow (result *from*) sophisticated, preconscious analysis.

In sum, Velmans (1991) distinguishes three senses in which information processing can be considered 'conscious': (a) when one is conscious of the process (e.g., the introspective aspects of planning and thinking); (b) when one is conscious of the results of the process, i.e., focal-attentive processing accompanied by consciousness (e.g., input analysis); and (c) when consciousness in some sense causally influences or enters into the process. Velmans' article is an attempt to show there is no human information processing that counts as type (c).

---

[17] Velmans is concerned with consciousness$_p$ here: "In the analysis that follows, it is 'consciousness' in the sense of 'awareness' that is of primary concern." (Velmans, 1991, p.651).

If Velmans' thesis is correct, there are serious repercussions for information processing models that propose a functional role for consciousness$_p$. Take, for example, one such proposed general function: the information that 'enters' consciousness (becomes conscious) is that which is made available to the cognitive system as a whole (e.g., Baars', 1988, global workspace model; see chapter 3). On Velmans' view, however, this function cannot be a function of consciousness$_p$, for awareness results from focal-attentive processing, and does not enter into it. If anything is going to function so as to make certain information available to other systems, it will be the focal-attentive processes themselves, not consciousness$_p$.

All cognitive or information-processing models appear to exclude phenomenal experience from their workings, even those that propose a functional role for it (for that functional role can be equally filled by some nonconscious$_p$ process). These models are concerned solely with the transformations of input to behavioural output, and therefore consciousness$_p$ seems inessential. Despite his attachment to this overall approach to studying the mind, Velmans sees a place for consciousness$_p$ in our accounts of human cognition. On his view, consciousness$_p$ can be seen from the third-person perspective to feature prominently alongside much information processing. For example, the contents of consciousness can often be regarded as 'output' accompanying certain information processing (as demonstrated in countless psychological experiments), and there seem to be significant functional differences between the processes that are accompanied by consciousness and the processes that are not. But from such a third-person perspective, awareness is dissociated from cerebral processing, and therefore appears 'epiphenomenal'. From the first-person perspective, however, consciousness$_p$ is not

epiphenomenal.

Velmans regards the first-person perspective as *"how things appear from the subject's point of view"* (1991, p.715, original emphasis). In other words, Velmans' reading of the first-person perspective closely resembles Nagel's 'what it is like to be X'. Although it is unclear whether Velmans would come out in support of properties of experience like qualia, he nevertheless considers there to be a point of view of experience that no third-person perspective of science can capture. Like Nagel and Jackson, he argues "that information processing models that view humans *only* from a third-person perspective are incomplete." (Velmans, 1991, p.716, original emphasis); and like Nagel's and Jackson's conclusion, Velmans' verdict is susceptible to the criticism that it is simply too early to tell for sure whether third-person science is or is not able to fully explain the first-person perspective. Still, unlike Nagel and Jackson, Velmans is not, strictly speaking, a new mysterian — he does not come to the pessimistic conclusion that no complete account of consciousness can be arrived at. Velmans remains optimistic about the prospects for explanations of consciousness, but he maintains that this cannot be achieved from the third-person perspectives alone. Furthermore, Velmans considers his article as showing "that first-person accounts can be *translated into* third-person accounts, but they cannot be *reduced* to them." (1991, p.717, original emphasis).

What reasons does Velmans offer for claiming that third-person science is incomplete, and for promoting the "complementary" principle — a marriage of the first-person and third-person accounts? A primary reason underpinning his view appears to be the intuitive implausibility of consciousness$_p$ being epiphenomenal. Relatedly, it seems that Velmans' insistence on the "complementary" principle is motivated by the apparent gap between scientific accounts of mind on the one hand, and first-personal and folk-psychological

accounts of mind on the other, leading to the intuition that the former will never be able to account for the latter. In line with what I have argued in previous sections, I will argue that unsupported intuitions do not provide a sound structure upon which to build one's theory of mind — one's intuitions can be wrong, or may lead us away from the real heart of the matter.

Velmans concludes from his survey of the third-person information-processing accounts that consciousness$_p$ appears epiphenomenal. However, this conclusion is not decisive, and his argument for it is unsound. As Block (1991) points out, Velmans' argument is faulty: just because consciousness$_p$ is not required for information processing does not logically entail that it does not actually enter into it. If one supposes that consciousness is dependent on the activity of a central executive system, then it could just as easily be argued that all the cases of information processing without consciousness$_p$ that Velmans surveys are not central executive functions, but rather functions of specialized modules. (A point with which Velmans would seem to agree, according to Block, 1991, p.670). If this is the case, then Velmans is right in claiming that the processes he surveys do not require consciousness$_p$, for they appear to proceed without any input from the central executive. Given this, however, Velmans is therefore wrong to claim that consciousness$_p$ does not have an effect on *all* information reaching the central executive. Maybe there are other cases of information processing whose execution does depend on certain information 'entering awareness'. Indeed, Block argues that Velmans actually provides evidence *against* his conclusion that consciousness$_p$ is, or appears to be, epiphenomenal.

There are further problems with Velmans' conclusion, however — criticisms that are not tied to explanations of consciousness that presuppose a

central executive system. (For as was argued in the previous chapter, there are some good reasons to distrust such explanations.)

Velmans is forced to say that from one perspective, $consciousness_p$ appears to be an epiphenomenon, and from another, it appears not to be an epiphenomenon. He then claims that it is the task of any complete theory of mind to resolve this paradox. But a complete theory of mind, for Velmans, is one that includes both third-person and first-person accounts. So just how is the matter to be adjudicated? Surely there must be a fact of the matter: either $consciousness_p$ (of whatever form) is epiphenomenal (in some sense of the term), or it is not. If $consciousness_p$ has causal status from the first-person perspective, then this fact can be explained, at least in principle, from the third-person perspective. Even if it turned out that this apparent causality, or at least its exact nature, is an illusion, that would not negate the importance or status of first-person accounts. As Velmans says, first-person accounts cannot be *reduced to* (explained away by) third-person accounts. However, *pace* Velmans, the third-person perspective *does* hold greater currency *if* its theories *explain* just why it is that the first-person perspective provides the sort of view that it does (viz. $consciousness_p$ having causal efficacy).

It appears that Velmans fails to distinguish between explanations of *general* psychological properties, capacities, states, or events, and explanations of *specific* (individual) psychological properties, capacities, states, or events. The former are concerned with questions like: how does the brain represents the world?; what are emotions?; and, of course, what is consciousness? The latter, on the other hand, are concerned with questions about individual thoughts, emotions, intentions, intuitions, actions, and the like. The proper business of psychology is answering questions of the general type, not the specific

questions, applicable only to particular individuals at specific times and in specific circumstances (see e.g., Wilkes, 1978).

In light of this distinction, then, we can see that it is not possible to 'translate' *individual* first-person accounts into third-person accounts, for as Velmans acknowledges, neither perspective can be subsumed under the other. But this is no challenge to third-person science, for science is not in the business of explaining the complete nature of mental phenomena as they appear to any one individual. If there is any 'translation' of first-person perspectives into third-person science, it will be of some general explanatory nature — for example, that epiphenomenalism appears not to be true from the first-person perspective because *this* is what usually occurs in the brain, and *this* is how the brain-world interaction operates, and *this* is how humans function as social beings, and so forth. Such a third-person account *explains* why things seem as they do, in a broad sense, but it does not explain away that first-person view, as Velmans seems to imply.

Velmans wants the third-person and first-person accounts to be given equal consideration, saying that "[a] complete psychology requires both." (1991, p.667). Moreover, he says that "[w]hether first-person or third-person accounts are more *useful*, depends entirely on the explanatory context." (1991, p.715). I couldn't agree more: folk-psychological explanations and predictions of everyday actions (in terms of first-person perspectives, and third-person accounts of first-person perspectives) are extremely useful in everyday human interaction, and, *pace* Paul M. Churchland (e.g., 1985, 1990), there is little reason to suppose that they will one day be *replaced* by some scientific third-person account. Likewise, scientific-psychological explanations and predictions are very useful (nay, crucial) in scientific explanatory contexts. But if scientific

psychology is to hold any currency at all, it must be able to explain *why* folk explain and predict their own and others' dispositions and actions in the way that they do, and *how* they do so.

These points are all in line with what has been argued in previous sections. Drawing explanatory links between the first-person perspective (viz. subjective reports of conscious experience) and brain properties is permissible, when we can establish they are reliably linked, because in such cases we have a prior commitment to the existence of conscious experience. Moreover, such inferences to the best explanation do not require us to *reduce* (identify) the introspective reports of the contents of consciousness to properties of the brain. Both consciousness_p and the neurophysiological and computational correlates of consciousness (consciousness_c) are on the table to be explained by science, and no a priori argument — especially one dependent upon intuition — will succeed in establishing that consciousness_p cannot be so explained. We will just have to wait and see what future scientific theories have to say, from which we will then have a better idea of how to judge ontological questions about consciousness. (And as will become evident in later sections of this and the following chapter, some theories are beginning to shed some light on this issue.)

What then of the claim that consciousness is epiphenomenal from the third-person perspective of science? At one point Velmans says, "consciousness is a form of *output* (associated with focal-attentive processing) that does not enter into cerebral processing." (1991, p.667, original emphasis). In the very next sentence he claims that this conclusion "appears to support epiphenomenalism (the view that brain events have causal effects both on other brain events and conscious experiences, but conscious experiences have no causal effects on the brain ...)" (p.667). The two interpretations of his conclusion do not amount to

the same thing. The latter claim is much stronger: not only does consciousness$_p$ not enter into (causally influence) information processing, it has *no causal effects on the brain whatsoever*. Clearly this form of epiphenomenalism is *too* strong (to the point of being untenable) for a physicalist to accept, as Dennett (1991a) has ably demonstrated (see chapter 3, above). If consciousness$_p$ exists, then it must have *some* effect in the physical world. There are two possibilities: either consciousness$_p$ consists in some form of neural activity; or it is the by product of some neural activity. If the second possibility is the case, then consciousness$_p$ would show itself as an epiphenomenon in some relatively innocuous (but hard to believe) way — maybe in terms of brain colour, or the heat generated by cerebral activity, as Dennett (1991a, p.405) humorously suggests. This may be a defendable (if not slightly outlandish) position, and it is certainly no challenge to physicalism. However, this can't be the sense of epiphenomenalism that Velmans has in mind, for he states that his conclusion is not "consistent with physicalism (the view that consciousness is ontologically identical to a physical state of the brain) — unless one is willing to accept that some cerebral states exist that neither have a function in themselves nor influence the development of subsequent, functional, cerebral states or processes." (Velmans, 1991, p.666).

Velmans appears not to want to accept the possibility of functionally and causally inert neural activity, however (although he is not entirely clear on this point), and therefore he must concede that the first possibility — that consciousness$_p$ could consist in some form of neural activity — is false. So where does this leave Velmans? He insists that consciousness exists as "a form of *output* accompanying certain forms of information processing" (Velmans, 1991, p.666, original emphasis). But what exactly does he mean by 'output'? What form does it take, and how and why are these brain states, or their

physical properties, epiphenomenal? As has just been ascertained, neither possible interpretation of the physical form that this output could take is tenable from Velmans' position. Moreover, echoing the worries over Cartesian materialism discussed in the previous chapter, output to where?

Even if consciousness$_p$ could be regarded as 'output' of some sort, why would it exist if it does, or did not, have any causal role to play in the cognitive-behavioural economy? If consciousness$_p$ is neither necessary for cerebral processing, nor actually enters into it (as Velmans would have it), then couldn't we just as easily get along without it? That is, would there be no noticeable behavioural differences without consciousness$_p$? Velmans cannot claim that there would be a difference if we did not have consciousness$_p$, for then he would have to give up his conclusion that consciousness$_p$ has no causal effect in the cognitive-behavioural economy. Therefore, Velmans is left in the unenviable position of having to proclaim that 'zombies' (nonconscious$_p$ persons behaviourally indistinguishable from conscious$_p$ persons) are possible; and as Dennett (1991a) has cogently argued, either zombies are impossible, or we are all 'zombies'!

In sum, if science proclaims consciousness (of whatever form) to be epiphenomenal, then so be it. But there better be some very good reasons for making such claims, for they are contrary to our first-personal intuitions. This is not to say that our first-personal intuitions will hold sway no matter what, however, for if scientific theories do proclaim consciousness to be epiphenomenal, then they must also eventually explain why it appears not to be epiphenomenal from the first-person perspective.

4. *Conclusion.*

There are no such properties as qualia. Therefore 'in principle' or *a priori* objections centred on qualia do not stand in the way of *S-F* theory and other forms of physicalism in their attempts to explain (whatever it is we mean by) consciousness. Of course this does not rule out the possibility that consciousness$_p$ exists in some form or other. Indeed *because* phenomenal consciousness certainly exists from the first-person perspective, as good scientists we are compelled to seek explanations of just what consciousness$_p$ is, as well as explanations of consciousness$_c$. There appear to be no significant impediments to explaining the various facets of consciousness$_c$, but consciousness$_p$ poses some more serious problems. Some inroads into our understanding of consciousness$_p$ and its possible neural implementation have been made, but these theories are still in their infancy. Consequently the gap between the first-person and third-person perspectives on consciousness remains open. One vexing facet of this gap between the perspectives is the issue of the apparent epiphenomenalism of consciousness$_p$. Although this issue remains unresolved, no *a priori* argument grounded in intuitions can hope to settle the matter once and for all. Scientific theory holds our only real hope of resolving the debate.

CHAPTER 5.

DRAWING THE THREADS TOGETHER: THE NEED FOR AN INTERDISCIPLINARY APPROACH.

1. *Consciousness and natural kinds.*

'Consciousness' is a folk-psychological, 'second order' concept and an everyday mental term (chapter 1). The EMTs are a heuristic overlay upon the behaviour of an embodied being acting in various environmental and social contexts. We use mental terms to describe and explain the states and behaviours of ourselves and others (Dennett's, 1978a, 1987, "intentional stance"). But do some EMTs refer to (even roughly) specific types of physiological or structural-functional states of the brain? If so, is 'consciousness' one of these terms?

The general folk-psychological concept of consciousness is not a suitable term for adoption and adaption by science (chapter 1). But are there any features or properties inherent in the use of the term that are suitable candidates for scientific constructs? Science is making reasonable progress in constructing theories of consciousness$_c$ — the information processing and neurophysiological mechanisms and processes responsible for the various phenomena of consciousness. But there are problems with the notion of consciousness$_p$ — it just seems as if there is no place for such a concept in physicalist science. Is consciousness$_p$ a suitable concept for adoption and adaption by science? Does consciousness$_p$ approximate a natural kind, or is it a concept of the folk-psychological, first-personal world, with no place in a scientific taxonomy?

Qualia do not exist, and therefore appeals to such properties do not erect impassable barriers to science explaining (or explaining away) consciousness$_p$. No amount of *a priori* argument will establish that physicalism cannot explain the first-person nature of experience. Dennett (1978d) puts this point rather nicely when he says:

> "If an empirical psychological theory develops that is both strongly confirmed and predictive of the rich variety of phenomena of consciousness, we can inspect it for an answer to the question ['Does functionalism/ physicalism leave something out?']. If it contains a theoretical role for something like qualia, we shall 'countenance' qualia in our ontology, but as theoretical entities, not epiphenomena; if no such role appears to be filled, then the very power of the theory will undermine the intuitions that now make the denial of qualia so counterintuitive." (p.256).

Questions of the existence and nature of consciousness$_p$ are empirical issues, to be decided by our best physicalist theories. We will just have to wait and see whether science comes up with the goods. What then is the best way for science to proceed so as to achieve this goal of explaining consciousness$_p$?

2. *Levels of nature and levels of explanation.*

In chapter 2 I introduced the idea that nature is multi-levelled, and that our theories of complex natural-world phenomena, especially the mind-brain, must reflect this multi-levelled structure. Multiple levels of nature require multiple levels of description and explanation. Some theorists claim that full-blooded theories of mind-brain and psychological phenomena must themselves be constructed in a multi-levelled fashion (e.g., Dennett, e.g., 1978a, 1987; Marr, 1982; Sterelny, 1990). These theorists, using different terminologies, claim that there are three principle theoretical domains or levels in psychology. Sterelny (1990) calls these levels the ecological, the computational, and the level of

physical implementation. (I adopt Sterelny's terminology here, because I find it the most informative and least confusing, especially with respect to the ecological level. As Sterelny points out, the other theorists' characterizations of this 'uppermost' level are sometimes misleading.)

The ecological level is the level at which cognitive capacities are broadly characterized in terms of what a system can do, while remaining neutral on mechanism. An ecological characterization of a cognitive capacity specifies a system's overall information processing competence, and therefore "provides a precise understanding of the information extraction problem the system solves." (Sterelny, 1990, p.45). A computational level theory details a method by which the system of interest might perform the informational function described at the ecological level. Theories of this type specify how a cognitive mechanism performs a task in precise computational-functional terms, ultimately in the form of algorithms. Computational processes must be physically realized; the task of theories at the level of physical implementation is to detail the relevant physical structures and processes of particular systems, showing how they perform the said computational processes.

The individual structures and processes that constitute psychological phenomena are likely to be members of many natural kinds. The above division of explanatory levels appears to be an effective and reasonably well agreed upon demarcation of the domains in which the multiple natural kind categories of psychological phenomena are to be found. However, notice how Lycan's exposition of the multiple-level view (chapter 2) blurs the distinction between the computational level and the level of physical implementation; we can talk of both computational function and physical implementation at (virtually) all levels of analysis. Hence there is no sharp distinction between cognitive psychology and neuroscience; a view with which I concur.

Note also that applying the three levels of analysis to 'consciousness' is a "risky oversimplification", as Dennett (1991a, p.277) warns, for doing so may blind us to the possibility that some features of consciousness may have multiple functions, or may be rather poor at achieving these functions, or may have no functions at all. Indeed, given that 'consciousness' denotes a disparate and heterogenous assortment of cognitive, experiential, and behavioural phenomena, it would be a mistake to restrict the construction of one's theory of consciousness to the design specified by this hierarchy of theoretical domains.

Not all psychological phenomena lend themselves to this three-tiered analysis. Even if all the individual cognitive functions and processes associated with consciousness can be formulated by way of the three levels, consciousness$_p$ cannot be so analyzed; we are still left with the question of whether consciousness$_p$ is a natural kind. Phenomenal experience — what it is like to perceive X, be a bat, etc. — is seen not as a capacity, function, or process, but as a property or set of properties, about which we have no idea how (or even whether) cognitive or brain processes give rise to them.

If there is no sharp distinction between cognitive psychology and neuroscience, between function and implementing structure, then theories of mind-brain are structural-functional (*S-F*) theories (chapter 2). *S-F* theory is paradigmatically multidisciplinary: Patricia S. Churchland (1986), for example, characterizes neuroscience as including clinical neurology, neuropsychology, and cognitive neurobiology, and she advocates "unified" theories of mind-brain, the development of which will see the theories of "the assorted psychological sciences" (p.9) and ethology "co-evolve" with neurobiological theories.

Of particular importance to interdisciplinary approaches to explaining the mind-brain and mental phenomena is the issue of intertheoretic reduction. Will psychological theories be reduced to neuroscientific theories, therefore eventually eliminating the need for psychology? Or are one or more of the psychological sciences autonomous? While lack of space precludes a detailed examination of intertheoretic reduction, my adherence to the multiple level view, with its implications for theories explaining the phenomena of consciousness, leads me to the threshold of this thorny polemic.

3. *The craft of folk psychology.*

If, as was put forward in chapter 1, folk psychology is understood primarily as a craft rather than as a primitive version of a systematic theory of behaviour and mental life, then the study of the craft is not a candidate for intertheoretic reduction. (Although if one wishes to take up the 'theory' theory of folk psychology, *in addition to* the craft theory, then it is possible that this 'theory' theory is a candidate for reduction or replacement, as the Churchlands argue.) Even if an advanced neuroscience replaces or eliminates all folk-psychological concepts, there will be a place for a theory of the craft of folk psychology. A theory of the craft of folk psychology, in contrast to the 'theory' theory, is in the business of explaining how and why people use this system (*any* system) of explanatory and ascriptive concepts, and how and why their use evolved; it is not in the business of unearthing the real or tangible referents (states, events, processes) of these terms. Folk psychology, understood as a craft, is not a crude attempt at explaining the structures and processes of the mind-brain, but an attempt at explaining the social actions and interactions of people.

Notice that the craft theory of folk psychology fits rather nicely with Dennett's (e.g., 1978a, 1987) instrumentalist stance on intentionality. The craft view of folk psychology is a perfect bedfellow for the view that our common sense practice of explaining behaviour is *merely* a fortuitous non-factual overlay or level of description, i.e., the view that *"all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy" (Dennett, 1987, p.29). However, in accepting the legitimacy and relative autonomy of the study of the craft of folk psychology, one is not required to take up such a strong instrumentalist stance. Lyons (1990), for instance, argues that the intentional stance *is* "a factual overlay about our internal psychological economy yet not an overlay of facts about our neurophysiology." (p.263). He suggests that it is unlikely folk psychology as we know it would have evolved if it were just a pragmatic, non-factual overlay upon the behaviour of an embodied being acting in the world. Folk psychology proposes a host of intervening variables between perceptual input and behavioural output, and therefore presupposes that humans are information processing systems. Lyons argues that folk psychology is right in considering humans as information-processing devices, even if it is wrong about how we operate as such devices. The intervening variables posited by folk psychology are supposed to be in-the-head states, but they are so only derivatively; the mental states of folk psychology are in the first instance linked to publically observable facts about the environment and behaviour. On this view, folk-psychological talk is likely to pick out real, if inexact and incomplete, patterns in our behaviour, and derivatively, in our brain states. Thus the concepts supplied by folk psychology are not natural kinds, although they may point to them (sometimes very roughly), and are more likely to be replaced or eliminated than

reduced by an advanced neuroscience. Nevertheless, folk psychology may tell an approximate story — a story that needs to be examined for what it is: a craft or practice[18], predicated on the assumption that the content of mental states is supplied by the physical and social environment.

Thus both the strong instrumentalist line and the position as advocated by Lyons (1990) are consistent with the claim that the study of folk psychology is a legitimate enterprise even when a mature neuroscience arrives on the scene. Viewing folk psychology as a craft reveals two important implications for our theories of consciousness: (1) *S-F* theory and the rest of natural science may not be able to offer comprehensive explanations of all that is implied by folk psychology's 'consciousness'. Comprehensive theories of the phenomena of consciousness will require a rapprochement of the natural and social sciences. (Section 4 below.) (2) Studying the craft of folk psychology is likely to contribute significantly to our theories of the phenomena of consciousness, for 'consciousness' is a folk-psychological concept, bound up with the theory of mind implicit in the craft. (Section 5 below.)

4. *The missing component: social constructs.*

4.1 *What is missing?*

Presumably there are brain mechanisms and processes responsible for our ability to perform the craft of folk psychology. So the cognitive and neuro-sciences will have a large role to play in the study of this skill. I agree with the Churchlands (e.g., Churchland and Churchland, 1983; P. S. Churchland, 1986; P. M. Churchland, 1988) that neuroscience has a crucial role to play in

[18] Note that this claim is an extension of Lyons' (1990) position; I do not know if he would endorse it.

psychological explanation, especially in explaining how the brain can represent the world. However, I contend that folk psychology, understood as a craft, cannot be subsumed under some future neuroscientific theory, because a crucial element would be left out. This is *not* to concede defeat to the new mysterians, however, for what I see to be missing from many a physicalist's picture is not some unexplainable intrinsic feature of experience, but an examination of the fundamentally social nature of the 'human condition', and hence of social structures and other social constructs.

Neuroscience is concerned with the sub-personal aspects of human functioning, and can be considered to incorporate the computational level in addition to the level of physical implementation (see section 2 of the present chapter). Social structures and the social milieu, on the other hand, are paradigmatically at the personal and collective levels of description. Is it not possible that *both* the subpersonal and the personal views, the natural and the social sciences, are required for full accounts of mind and behaviour?

Certainly I do not believe that the personal level can be reduced to the sub-personal level, although some developed personal-level theory may be *explained* by some developed subpersonal-level theory (this is how, for example, Patricia S. Churchland, 1986, interprets 'intertheoretic reduction'). It may be, as Van Gulick (1980) trys to show, that a broadly conceived functionalist account of mind can be extended to accommodate the social dimension. But even if this is the case, we must bear in mind that this broadly conceived functionalist account should not be limited to any one particular explanatory level. For the major problem with restricting oneself to a limited number of levels of description and explanation is that useful and valid generalizations obtainable at other levels will be omitted. (A compelling

illustration of the loss of explanatory power by restriction to lower levels of description and explanation is Putnam's, 1975, famous 'square peg in round hole' thought experiment. Given only a microstructural description of a square peg and a round hole, of equal cross-sectional area, one would not readily see, if at all, that the peg will not fit into the hole.)

4.2 *The self as a social construct.*

A look at some of Mead's (e.g., 1964; see Oatley, 1988) views on consciousness is somewhat revealing of the likely inadequacies of a purely neuroscientific, or analytical, structural-functional account. Mead held that the existence of the conscious mind cannot be explained simply as a higher order property emergent from complex brain functioning. He hypothesized that full human consciousness requires comparisons with a socially derived sense of self[19]. Consciousness, mind, and the self arise in relation to the close and often complex social interactions we partake in throughout our lives. (Others to make a case for the social construction of the sense of self and, therefore, for consciousness being socially derived, include Humphrey, 1983, 1986; and Marcel, 1988.)

In the discussion of Johnson-Laird's (1983, 1988a, 1988b) theory of consciousness (chapter 3), I introduced the concept of *mental models*. The idea of mental models dates back to Craik's (1943) hypothesis that the human brain

---

[19] Note that this notion of the social self is to be distinguished from the 'biological self', which refers to an organism's sensitivity to the boundaries of itself as an organism in relation to the outside environment. This latter distinction between the self and nonself is served by a functional split in the nervous system – even the nervous systems of the simplest of organisms will be segregated into "equipment subserving, on the one hand, internal hedonic regulation and, on the other hand, information processing about the state of the external world." (Flanagan, 1990, p.325). See also e.g., Dennett (1991a); Edelman (1989).

is capable of modelling aspects of the outside world, and of itself. The brain can be viewed as an "anticipation machine" (Dennett, 1991a) because it can track key features of the environment and anticipate future events by 'running simulations' of actual and possible events, including the effects of possible actions based on previous experience. Mead proposed that human consciousness requires a socially derived model of self, and Johnson-Laird formulates this proposal in the modern garb of cognitive science, in the form of Craikian models. (See also the discussion in Oatley, 1988.)

Dennett (1991a) calls the socially derived self the "center of narrative gravity", for two reasons: he sees the construction of a model of the self as akin to telling a story or weaving a narrative, i.e., it is a process of representation (to ourselves and others) via language and gesture; and he regards the self not as a concrete entity, but as an artifact of social processes, and hence as one of a set of *abstracta* (logical constructs), analogous to the physicist's centre of gravity. (Incidentally, on Dennett's instrumentalist stance, beliefs are also *abstracta*.)

4.3 *But can this shed any light on consciousness$_p$?*

Oatley (1988) points out that theories of the likes of Johnson-Laird's — those proposing recursive model construction of the self — tell us only about the mechanisms by which this modelling occurs, and not about phenomenology (consciousness$_p$). He claims that theories about the *content* of mind will also be required if we are to achieve anything approaching complete explanations of the phenomena of consciousness. What is needed, according to Oatley, is a rapprochement between natural science and social science, for "the phenomenology of explicitly knowing, and knowing that we know, derives from the *socially* derived experience of the sense of self as director and as part

of the comparison processes of consciousness." (p.378).

While I have little doubt that *some* rapprochement between natural and social science will go a long way to explaining the intricacies of $consciousness_p$, it is not clear from Oatley's brief exposition that his proposal will rid $consciousness_p$ of all its mystery. That is, without a more fully developed proposal, it is too early to tell whether the intuitive gulf between third-person explanations of the phenomena of consciousness and the first-person 'feel' of experience will be bridged, or even significantly diminished. Fortunately there is such a proposal in the offing: Dennett's (1991a) 'Multiple Drafts' model of consciousness.

4.4. *Dennett's proposals.*

One important feature of the social dimension that, if it is to have any explanatory bite, cannot be captured at the microstructural level of neuroscientific theory, is the "meme"[20]. While genes are the units of biological evolution, memes are the units of sociocultural evolution. Moreover, memetic evolution is a process which obeys the laws of natural selection, just as genetic evolution does. Memes can come in many forms, but are most generally known as ideas or units of knowledge. The notion of meme evolution plays an important part in Dennett's (1991a) theory. The infestation of the human brain by hordes of memes produces considerable change in the proficiency of that organ. "The haven all memes depend on reaching is the human mind, but a human mind is itself an artifact created when memes restructure a human brain in order to make it a better habitat for memes." (Dennett, 1991a, p.207). Furthermore, the functional differences created in brains by memes, and observable in behaviour, "though presumably all physically embodied in

[20] The term was coined by Dawkins (1976).

patterns of microscopic changes in the brain, are as good as invisible to neuroscientists, now and probably forever, so if we are going to get any grip on the functional architecture *created* by such meme infestations, we will have to find a higher level at which to describe it." (Dennett, 1991a, p.210).

What is this level of description? Dennett suggests borrowing once again from the language of computer scientists: the level of description and explanation at which the effects of memes are evident is analogous to a 'software level' of description. In particular, Dennett claims that consciousness can be understood in terms of the operation of a 'virtual machine' created by the effects of large numbers of memes in the brain. In computer science, a virtual machine is a higher-level programming language, i.e., a systematized set of rules and instructions. Programming languages allow a user to interact in a particular way with a computer, and determine what sort of functions or jobs that computer can perform. "So a virtual machine is a temporary set of highly structured regularities imposed on the underlying hardware by a *program*: a structured recipe of hundreds of thousands of instructions that give the hardware a huge, interlocking set of habits or dispositions-to-react." (Dennett, 1991a, p.216).

A 'user illusion' is created by a virtual machine: it gives the operation of the hardware upon which it is implemented a characteristic style and appearance to the user (e.g., my Macintosh may at one moment be operating as a Write Now word processor, and at another moment as a Hypercard filing system). The particular user illusion created by memes in the brain is "von Neumannesque", i.e., it is of a system operating serially, controlled by a central processing unit. "Conscious human minds are more-or-less serial virtual machines implemented — inefficiently — on the parallel hardware that

evolution has provided for us." (Dennett, 1991a, p.218). The 'user' in this case consists in the interacting systems that have some degree of access to the output of other systems (maybe via the global workspace), and that together form the neural basis of the centre of narrative gravity.

In conclusion, social theory is quite compatible with physicalism. There are no physicalist strictures upon the general concerns of the social sciences (except when those sciences posit entities that are not based entirely within the physical world). However, the compatibility of levels of description and explanation need not imply a reduction or elimination of the higher levels by our best lower level theories. As Sterelny (1990) concludes, "[p]hysicalism does not require that every good theory be reducible, ultimately, to physical theory. So the fact that folk psychology, and much of cognitive psychology, is unlikely to reduce to the neurosciences does not commit us to their 'reconfiguration' or elimination." (pp.205-206). Major headway can be made in the development of theories of mind and consciousness if we retain a sociocultural level of description and explanation. (Hence the value of anthropology to cognitive science; see e.g., D'Andrade, 1981.) For example, the concepts of memetic evolution and the socially constructed self cannot be captured by neuroscience. I now turn to a discussion of another major advantage in retaining a sociocultural explanatory level: how a study of the development and practice of the craft of folk psychology can contribute to our theories of the phenomena of consciousness.

5. *The evolutionary development of consciousness, and the connection with our competence for the craft of folk psychology.*

Recall that a close parallel can be drawn between folk physics and folk psychology. On this view, if the concepts and 'laws' of folk physics allow us to assess and predict the physical world, then the concepts and 'laws' of folk psychology constitute our ability to assess and predict our social world. Folk-psychological principles are first and foremost items of the social realm; they may also, by implication, be about brain states and events, but that is not their *raison d'etre.* Just as a roughly accurate grasp of some basic physical principles of an organism's environment, relative to its needs and sensory and motor endowments, is vital for the survival of that organism, so too might a roughly accurate grasp of some basic psychological principles be vital to the survival of a social organism.

Humphrey (1983) argues that it was sociocultural evolution, running parallel to, and parasitic on, biological evolution, that was primarily responsible for the extraordinary advances in the development of the human mind. Humphrey's 'just so' story of the development of consciousness runs like this: As our ancestors became increasingly dependent on social interactions and groupings, an advantage was accrued to those who had the capacity to explain and predict their conspecific's behaviour. Initially the explanatory system would have been rather crude: these early hominid ancestors lacked the capacity for 'introspection', but they could observe environmental conditions and behaviour, 'what went in and what went out' (i.e., they implicitly treated fellow humans as 'information processing devices'), and so could "have pieced together an external, objectively based explanatory model" (Humphrey, 1983,

p.50). Once social interactions became more complex, the possession of a more complex explanatory system would have proved a major advantage. Humphrey argues that a new method of psychological understanding was selected for, one that depends on the human ability to 'introspect'. By examining the contents of her own consciousness, an individual reasons by analogy to develop a model of the behaviour of others.

But what does Humphrey mean by 'introspection' and 'the contents of one's own consciousness'? Introspection he sees as the capacity for 'reflexive consciousness', i.e., to be aware of being aware (conscious of being conscious). Humans have developed the capacity of being consciously aware of some of the results of the workings of their own brains. That is, the information processing of the brain produces certain motivational and dispositional states, which are accompanied by *feelings* (sensations, emotions, volitions, etc.) only when one is consciously aware that one's brain is in the relevant states. So our early ancestors' "brains would receive and process information from their sense-organs without their minds being conscious of any accompanying sensation, their brains would be moved by, say, hunger or fear without their minds being conscious of any accompanying emotion, their brains would undertake voluntary actions without their minds being conscious of any accompanying volition" (Humphrey, 1983, pp.48-49). But once the higher-order capacity of reflexive consciousness evolved, voluntary actions could be accompanied by volitions, motivational states could be accompanied by feelings, and information received by the sense-organs could be accompanied by sensations.

Marcel (1988) proposes a similar thesis, but is more explicit about the importance of a socially constructed model of the self. He claims that "it is the operation of reflexive consciousness which creates phenomenal experience from

otherwise non-conscious data. To the extent that reflexive monitoring can only be directed on the basis of a model of what is to be monitored, which is surely partly socially constructed, our phenomenal experience is the realization of a social construct." (p.150).

The heart of Humphrey's theory is this: full human consciousness and the skill or craft of folk psychology (he calls it "natural psychology") are interdependent; reflexive consciousness, the capacity to gain a picture of the 'psychological structure' underlying one's own behaviour (including a concept of the self), provides the means by which a conceptual framework can develop as a method for explaining the behaviour of one's fellows. "Nature's solution to the problem of doing psychology has been to give every member of the human species both the power and the inclination *to use a privileged picture of his own self as a model for what it is like to be another person.*" (Humphrey, 1983, p.6, original emphasis). The conceptual frameworks that have developed are at a level of description that is the most pragmatic and useful for an explanatory system for human behaviour — not at the level of neurophysiology, but at what has proved to be an equally valid level: the psychological. Nevertheless, the extent to which folk psychology has proved to be successful "is presumably because the workings of my conscious mind do in reality correspond in some formal (if limited) way to the workings of my brain." (Humphrey, 1983, p.48; see also Lyons', 1990, similar point, mentioned in section 3, above).

Humphrey (1986) uses the analogy of an 'inner eye' to explain our ability to introspect; however he is careful not to use the term in any way other than metaphorically. Via brain processes of *some* sort, individuals gain knowledge about the 'psychological structure' underlying their own behaviour by having the relevant experiences.

Dennett (1991a) is more explicit on the possible means by which we come to 'know' about our brain states. According to Dennett's theory, the thesis that individuals gain a picture of their 'psychological structure' is only partially correct: the picture they get is an *illusion* of psychological structure, not a roughly accurate account of what really goes on in the brain. Rather than adopt the folk-psychological practice of individuating contentful mental states (e.g., a state of being aware that one is seeing red, or a thought about that state of awareness), science should look to the *processes* that are responsible for the correlation between the information processing events of the brain and an individual's capacity to report on the information in those events (this is the basis of Dennett's method of "heterophenomenology", as discussed in chapter 4). For on Dennett's view, it is the reporting of one's experience (e.g., 'I have the sensation of seeing red'), and not the lower-order states themselves (e.g., sensing red), that gives rise to there *seeming to be* contentful higher-order states.

That there 'seems to be' contentful higher-order states, or recursive levels of reflexive consciousness (see especially Rosenthal, 1986), is an illusion, according to Dennett (for remember, he is an instrumentalist) — a user illusion created by the virtual machines of the brain. The process of reporting to ourselves and to others about our experience "is precisely what creates or fixes the content of the higher order thought expressed" (p.315)[21]. Confabulation is the name of the game: we construct multiple 'narratives' or stories about our experiences (hence the 'Multiple Drafts' title to his theory). These reports about

---

[21] Karmiloff-Smith (e.g., 1986, 1991) provides intriguing accounts of the developmental stages children go through in constructing higher-level representations (e.g., in the context of problem solving), and how this is linked to linguistic competence. She argues that children are metatheoreticians: they discover how the physical and social worlds operate by constructing *theories*, not by simple observation of the facts.

experience do not originate from the outputs to the speech centre by an omniscient central homunculus or self, but from a Pandemonium-style process of (private and public) speech production (see chapter 3).

Dennett (1991a) concludes:

> "To put the point tautologically, since it really does seem to people that they have both these beliefs about their experiences, and (in addition) the experiences themselves, these experiences and beliefs-about-experiences are both part of how it seems to them. And so we have to explain that fact — not the fact that our minds are organized into hierarchies of higher-ordered representational 'states' of belief, meta-belief, and so-forth, but that our minds tend to seem to us to be so ordered." (p.319).[22]

And just what is this characteristic of 'seeming to be'? Why it would seem that it is our old friend, consciousness$_p$. But Dennett has sought to explain *away* phenomenology (especially when qualia are held to be its defining characteristics). He is a 'realist' about consciousness$_p$, but only in the sense that consciousness$_p$ is an *abstractum*, a logical construct, like beliefs and the self. (See Dennett and Kinsbourne, 1992, p.235. Dennett, 1991b, provides a more detailed case for this form of attenuated realism.) Dennett (1991a) concludes that there is no such thing as phenomenology, rather there just *"seems to be phenomenology"* (p.366). But as Strawson (1992) points out, this claim must be false, for 'seeming to be' *is* phenomenology: "Such seemings are phenomenological goings-on, and phenomenological goings-on are such seemings." (p.5).

In addition to this puzzle, there are further problems with the explanations of consciousness proposed by Dennett and Humphrey: namely, the reduction of

---

[22] It is this aspect of Dennett's theory that I think Block (forthcoming) misses when he claims that Dennett frequently appears to want a reduction of phenomenal consciousness to reflexive consciousness. Dennett (1991a) does not, *contra* Block, treat reflexive consciousness as a good analysis of consciousness for science; what he does claim is that it is a good analysis within *folk* psychology (see pp.306-320).

consciousness$_p$ to reflexive consciousness (e.g., Block, forthcoming), and the supposed intimate and necessary relationship between language and consciousness (e.g., P. S. Churchland, 1983). What should we say about the consciousness of pre-linguistic children and animals, for example? Part of the solution may lie in distinguishing self-consciousness from 'mere' or 'simple' consciousness, as Dennett suggests; although there is still some warranted scepticism over such proposals (P. S. Churchland, 1983, for example, doubts whether this distinction carves nature at her joints).

I do not have any ready solutions to these quandaries, except to say that Dennett's sketch of a theory provides possibly the best means yet of showing how to bridge the notorious explanatory gap between scientific theories of consciousness and the first-personal 'feel' of experience. The following section summarizes the usefulness of Dennett's theory in this regard.

6. *Metaphors and analogies: Keys to unlock the mysteries of mind.*

Large chunks of this thesis could be considered as constituting a favourable review of Dennett's (1991a) *Consciousness Explained*. I do not think he has explained consciousness, however, at least in any widely accepted, strong sense of 'explained'. Although he has, as Dennett himself admits, provided a *sketch* of a theory, which is a start: "My explanation of consciousness is far from complete. One might even say that it was just a beginning, but it *is* a beginning, because it breaks the spell of the enchanted circle of ideas that made explaining consciousness seem impossible." (p.455). Dennett's main concern in providing his theory sketch has therefore been to reduce many of the mysteries of consciousness and mind to puzzles; science cannot deal with mysteries, for they offer no clues as to how to begin solving

them (mysteries are, by definition, *inexplicable*). But once a mystery has been turned into a puzzle, it is possible for scientists to proceed in the directions stated or implied in the conversion (or indeed to change those directions, and hence the nature of the puzzle, should the weight of empirical evidence warrant it). (See Dennett's brief discussion of this point in Dennett, 1992.)

Dennett's primary tools for the task of turning mysteries into puzzles have been metaphors and analogies. In particular, he has found the software level of description to be the most productive source of metaphors for providing us with the blueprints of a bridge across the explanatory gap. "The concepts of computer science provide the crutches of imagination we need if we are to stumble across the *terra incognita* between our phenomenology as we know it by 'introspection' and our brains as science reveals them to us." (Dennett, 1991a, p.433).

The careful use of metaphors and analogies is an important initial step on the road to more complete and rigorous scientific theories. So how does this process reduce or bridge the explanatory gap? Or more generally, when does a mystery become a puzzle? A suggestion of the answers to these questions may be found in some proposals on the psychological nature of understanding. Although the concept of understanding is difficult to define, it undoubtedly involves knowledge, belief, and some degree of assumption on the part of the understander (Johnson-Laird, 1983). More specifically, individuals may be said to have a greater understanding of a phenomenon when the explanation they are offered achieves one or more of at least the following: (i) an increase in their knowledge of that phenomenon; (ii) a new way of thinking about that phenomenon, i.e., that provided by the explanation; (iii) an increase in the likelihood of their acceptance of (belief in) the explanation in question;

and (iv) the explanation places a minimal reliance on their intuitions, i.e., take little as possible for granted.

Certainly conditions (i) and (ii) may be met by the careful use of metaphors and analogies within nascent theories of mind-brain. Condition (iii) may just as easily be met by dualist or other supernatural theories, but if the understander is at least sympathetic to physicalist concerns, then the thoughtful use of metaphors and analogies may increase that person's acceptance of the physicalist theory offered. With regard to condition (iv), Johnson-Laird (1983) suggests that those explanations which can be put in the form of an effective procedure (see chapter 2) will have most explanatory value, because this will place minimal, or as near as possible to nonexistent, reliance on our theoretical intuitions. Thus Johnson-Laird proposes a computational constraint on the content of psychological theories; psychological theories will "count as putative explanations only if it is possible to formulate them as effective procedures — or at least those parts of them giving rise to empirical predictions." (1983, pp.7-8). Many of our current psychological theories are not ready for such precise formulation, however, for some of the phenomena they pertain to are, in Dennett's words, very much a part of the *terra incognita* between the first-person and third-person perspectives on the human condition. Although the use of metaphors and analogies tends to place a good deal of reliance on our intuitions, it is via these metaphors and analogies that scientists can point the way towards more rigorous formulations. Metaphors and analogies may be keys to unlock the doors labelled 'mysteries of mind', but behind those doors lie the labyrinthine corridors of puzzles.

7. *Where to now?*

One underlying general concern to emerge from this thesis is that anything approaching consummate explanations of all the neuroscientific, experiential, and behavioural phenomena underlying our use of the terms 'consciousness' and 'mind' will require a truly interdisciplinary approach. 'Consciousness' is a term of the vernacular that picks out such a disparate or heterogeneous group of these phenomena (chapter 1) that nothing short of a full scale attack from all relevant disciplines will uncover much of the mystery and confusion surrounding it. Thus I agree with Thagard (1989) when he says "I reject as *methodological imperialism* the opinion that other approaches are not worth pursuing as well [as the type of connectionist model he uses]. In the current neonatal state of cognitive science, restrictions on ways to study the mind are clearly premature." (pp.457-458). There is a place for all manner of descriptive levels in our explanatory quest. Our best general approach for most complex cases of psychological explanation will be integrative, without the requirement of strict intertheoretic reduction. Dennett (1991a), for instance, says of his integrative approach to a theory of consciousness: "The limited perspective of each enterprise [neuroscience and cognitive psychology/ AI] taken by itself just shows us the need for another enterprise — the one we are engaged in — that tries to put together as many as possible of the strengths of each." (p.256).

Dennett (1991a) claims that philosophers like himself are in an ideal position to develop such interdisciplinary theories (or at least sketches of theories, to be "made honest by modeling at [the computational] level"; p.268). The philosopher's and philosophical psychologist's job is to tie together the best theoretical claims from all manner of disciplines, while keeping a watchful eye

out for logical inconsistencies, conceptual problems, misguided assumptions, and barren claims (see also Rust, 1992). Moreover, before anything approaching 'complete' theories can be advanced, a detailed survey of the different sorts of questions that need answering, and how they might be answered, is required — a job ideally suited to philosophers and philosophical psychologists, but one that may be beyond the ken and skills of any one theorist (Dennett, 1991a).

If Dennett is right here, then we can see that the interdisciplinary approach is in dire need of systematic examination. Cognitive science is claimed to be interdisciplinary, yet it is not entirely clear exactly how the disciplines can and should relate to each other. Is there some formal program or methodology of scientific explanation that incorporates, or indeed champions, the interdisciplinary approach? If so, what does this philosophy of science have to say about how one should go about using and combining the theories and results of multiple disciplines? Is there some best 'middle way' between creating interdisciplinary theories that are too stringent and 'focused' in their postulates and objectives, and creating those that are too much of a 'mish-mash' of possibly unrelated, contradictory, or unsystematic claims and results? These are some of the questions which need to be addressed in a critical examination of the interdisciplinary approach to theories of cognitive systems, and especially of consciousness. In effect, what I am asking is whether there is, or can be, a metatheory of interdisciplinary explanation. If so, what form does or should it take? What bearing will the answers to these questions have on present and future explanations of the phenomena of consciousness? But these are questions for another thesis.

References.

Allport, A. (1988). What concept of consciousness? In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 159-182). Oxford: Clarendon Press.

Anderson, J. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Armstrong, D. M. (1968). *A materialist theory of mind.* New York: Humanities Press.

Armstrong, D. M. (1981). The causal theory of the mind. In *The nature of mind and other essays* . New York: Cornell University Press. Reprinted in W. G. Lycan (Ed.), (1990), *Mind and cognition: a reader.* (pp. 37-47). Oxford and Cambridge, MA: Basil Blackwell.

Armstrong, D. M. (1987). Mind-body problem: philosophical theories. In R. L. Gregory (Ed.), *The Oxford companion to the mind.* (pp.490-491). Oxford: Oxford University Press.

Atkinson, R. C., and Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 224,* 82-90.

Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge: Cambridge University Press.

Baddeley, A. D. (1986). *Working memory.* Oxford: Oxford University Press.

Baddeley, A. D. (1990). *Human memory: Theory and practice.* Hove, Sussex: Lawrence Erlbaum.

Baddeley, A. D. (1992). Consciousness and working memory. *Consciousness and Cognition, 1,* 3-6.

Baddeley, A. D. (unpublished). Working memory and conscious awareness.

Bechtel, W. (1991). A review of Boden, M. A. (1989). Artificial intelligence in psychology: inter-disciplinary essays. (Cambridge, MA: MIT Press/ Bradford Books.) *American Journal of Psychology, 104,* 279-310.

Bhaskar, R. (1975). *A realist theory of science.* Leeds: Leeds Books.

Billman, D., and Peterson, J. (1989). Critique of structural analysis in modeling cognition: a case study of Jackendoff's theory. *Philosophical Psychology, 2,* 283-296.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and cognition. Issues in the foundations of psychology, Minnesota studies in the philosophy of science, Vol. 9.* Minneapolis : University of Minnesota Press.

Block, N. (1980a). Are absent qualia impossible? *The Philosophical Review, 89,* 257-274.

Block, N. (1980b). Introduction: What is functionalism? In N. Block (Ed.), *Readings in philosophy of psychology, Vol. 1.* (pp.171-184). Cambridge, MA: Harvard University Press.

Block, N. (1990). The computer model of the mind. In D. N. Osherson and E. E. Smith (Eds.), *Thinking: An invitation to cognitive science. Vol. 3.* (pp.247-289). Cambridge, MA: MIT Press.

Block, N. (1991). Evidence against epiphenomenalism. Open peer commentary to Velmans, M. (1991), Is human information processing conscious? *Behavioral and Brain Sciences, 14,* 670-672.

Block, N. (forthcoming). Other things explained. Review of Dennett's 'Consciousness Explained'. *The Journal of Philosophy.*

Block, N., and Fodor, J. A. (1972). What psychological states are not. *Philosophical Review, 81,* 158-181.

Cam, P. (1988). Modularity, rationality, and higher cognition. *Philosophical Studies, 53,* 279-294.

Cam, P. (1989). Notes toward a faculty theory of cognitive consciousness. In P. Slezak and W. R. Albury (Eds.), *Computers, brains and minds: Essays in cognitive science.* (pp.167-191). Dordrecht: Kluwer Academic.

Cherniak, C. (1986). Limits for knowledge. *Philosophical Studies, 49,* 1-18.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78,* 67-90. Reprinted in W. G. Lycan (Ed.) (1990), *Mind and cognition: a reader.* (pp.206-223). Oxford and Cambridge, MA: Basil Blackwell.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy, 82,* 8-28.

Churchland, P. M. (1988). *Matter and consciousness: A contemporary introduction to the philosophy of mind (revised edition).* Cambridge, MA: MIT Press.

Churchland, P. M. (1990). Some reductive strategies in cognitive neurobiology (revised). In M. A. Boden (Ed.) (1990), *The philosophy of artificial intelligence.* (pp. 334-367). Oxford: Oxford University Press. Originally published in *Mind, 95,* 279-309.

Churchland, P. M., and Churchland, P. S. (1981). Functionalism, qualia, and intentionality. In J. I. Biro and R. W. Shahan, (Eds.), *Mind, brain, and function.* Oklahoma : University of Oklahoma Press. Originally published in *Philosophical Topics, 12,* 121-145.

Churchland, P. M., and Churchland, P. S. (1983). Stalking the wild epistemic engine. *Nous, 17,* 5-18. Reprinted in W. G. Lycan (Ed.) (1990), *Mind and cognition: a reader.* (pp.300-311). Oxford and Cambridge, MA: Basil Blackwell.

Churchland, P. S. (1983). Consciousness: The transmutation of a concept. *Pacific Philosophical Quarterly, 64,* 80-95.

Churchland, P. S. (1986). *Neurophilosophy.* Cambridge, MA: MIT Press.

Churchland, P. S. (1988). Reduction and the neurobiological basis of consciousness. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp.273-304). Oxford: Clarendon Press.

Churchland, P. S., and Sejnowski, T. J. (1989). Neural representation and neural computation. In L. Nadel, L. Cooper, P. Culicover, and R. M. Harnish, (Eds.), *Neural connections, mental computations.* Cambridge, MA: MIT Press. Reprinted in W. G. Lycan (Ed.) (1990), *Mind and cognition: a reader.* (pp.224-252). Oxford and Cambridge, MA: Basil Blackwell.

Clark, A. (1987). From folk-psychology to naive psychology. *Cognitive Science, 11,*139-154.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* Cambridge, MA: MIT Press/ Bradford Books.

Cole, D. J. (1990). Cognitive inquiry and the philosophy of mind. In D. J. Cole, J. H. Fetzer, and T. L. Rankin (Eds.), *Philosophy, mind, and cognitive inquiry.* (pp.1-46). Dordrecht: Kluwer Academic.

Copeland, B. J. (unpublished). *Minds, brains, and machines.* University of Canterbury cognitive science class manuscript, 1991.

Craik, K. (1943). *The nature of explanation.* Cambridge: Cambridge University Press.

Crick, F., and Koch, C. (1992). The problem of consciousness. *Scientific American,* (Sept.), 153-159.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72,* 741-765.

Cummins, R. (1983). *The nature of psychological explanation.* Cambridge, MA: MIT Press/ Bradford Books.

D'Andrade, R. G. (1981). The cultural part of cognition. *Cognitive Science, 5,* 179-195.

Davis, L. (1982). Functionalism and absent qualia. *Philosophical Studies,* 41, 231-251.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

Dennett, D. C. (1975). Why the law of effect will not go away. *Journal of the Theory of Social Behaviour,* 169-176. Reprinted in Dennett (1978a).

Dennett, D. C. (1978a). *Brainstorms: Philosophical essays on mind and psychology.* Montgomery, Vt : Bradford Books.

Dennett, D. C. (1978b). Toward a cognitive theory of consciousness. In Dennett (1978a), *Brainstorms: Philosophical essays on mind and psychology.* (pp. 149-173). Montgomery, Vt.: Bradford Books. Originally published in C. W.s Savage (Ed.) (1978), *Minnesota studies in the philosophy of science.* Minneapolis: University of Minnesota Press.

Dennett, D. C. (1978c). Two approaches to mental images. In Dennett (1978a), *Brainstorms: Philosophical essays on mind and psychology.* (pp. 174-189). Montgomery, Vt.: Bradford Books.

Dennett, D. C. (1978d). Current issues in the philosophy of mind. *American Philosophical Quarterly, 15,* 249-261. Reprinted in D. J. Cole, J. H. Fetzer, and T. L. Rankin (Eds.), *Philosophy, mind, and cognitive inquiry: Resources for understanding mental processes.* (pp. 49-72). Dordrecht: Kluwer Academic.

Dennett, D. C. (1982). How to study consciousness empirically, or nothing comes to mind. *Synthese, 53,* 159-180.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press/ Bradford Books.

Dennett, D. C. (1988). *Quining qualia.* In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 42-77). Oxford: Clarendon Press.

Dennett, D. C. (1991a). *Consciousness explained.* Boston: Little, Brown, and Co.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy, 89,* 27-51.

Dennett, D. C. (1991c). Two contrasts: folk craft versus folk science, and belief versus opinion. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science.*Cambridge: Cambridge University Press.

Dennett, D. C. (1992). In 'An interview with Dan Dennett'. *Cogito, 6,* 115-125.

Dennett, D. C., and Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences, 15,* 183-247.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition.* Cambridge, MA, and London: Harvard University Press.

Edelman, G. (1989). *The remembered present: A biological theory of consciousness.* New York: Basic Books.

Farrell, B. A. (1950). Experience. *Mind, 59,* 170-198.

Flanagan, O. (1990). Consciousness. In *The science of the mind (2nd ed.)* (pp.307-366). Cambridge, MA: MIT Press/ Bradford Books.

Fletcher, G. J. O. (1984). Psychology and common sense. *American Psychologist, 39*, 203-213.

Fletcher, G. J. O. (forthcoming). The scientific credibility of commonsense psychology. In K. Craik, R. Hogan, and R. Wolfe (Eds.), *50 years of personality psychology*. New York: Plenum Press.

Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy, 65*, 627-640.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1985). Fodor's guide to mental representation: The intelligent Auntie's vade-mecum. *Mind, 94*, 77-100. Reprinted in J. D. Greenwood (Ed.) (1991), *The future of folk psychology: Intentionality and cognitive science*. Cambridge: Cambridge University Press.

Fodor, J. A. (1987). Introduction: The persistence of the attitudes. In *Psychosemantics: The problem of meaning in the philosophy of mind*. (pp.1-26). Cambridge, MA: MIT Press/ Bradford Books.

Gazzaniga, M. S. (1988). Brain modularity: towards a philosophy of conscious experience. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science*. (pp. 218-238). Oxford: Clarendon Press.

Glymour, C. (1980). Good theories do. In A. P. Maslow, R. H. McKillip, and M. Thatcher (Eds.), *Construct validity in psychological measurement*. (pp.13-21). Princeton: Educational Testing Service.

Greenwood, J. D. (1991a). Introduction: Folk psychology and scientific psychology. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science*. (pp.1-21). Cambridge: Cambridge University Press.

Greenwood, J. D. (1991b). Reasons to believe. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science*.(pp.70-92). Cambridge: Cambridge University Press.

Haig, B. D. (1990). Psychology as philosophy, *N. Z. Journal of Psychology, 19*, 81-83.

Harman, G. (1989). Some philosophical issues in cognitive science: Qualia, intentionality, and the mind-body problem. In M. I. Posner (Ed.), *Foundations of cognitive science.* (pp. 831-848). Cambridge, MA and London: MIT Press/ Bradford Books.

Harré, R. (Ed.). (1986). *The social construction of emotions.* Oxford: Basil Blackwell.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* Cambridge, MA: MIT Press/ Bradford Books.

Hayes, P. (1979). The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the microelectronic age.* Edinburgh: Edinburgh University Press.

Hilgard, E. R. (1977). Controversies over consciousness and the rise of cognitive psychology. *Australian Psychologist, 12,* 7-26.

Hilgard, E. R. (1980). Consciousness in contemporary psychology. *Annual Review of Psychology, 31,* 1-26.

Hilgard, E. R. (1992). Divided consciousness and dissociation. *Consciousness and Cognition, 1,* 16-31.

Hofstadter, D., and Dennett, D. C. (Eds.). (1981). *The mind's I: fantasies and reflections on mind and soul.* New York: Basic Books.

Horgan, T., and Woodward, J. (1985). Folk psychology is here to stay. *The Philosophical Review, 94,* Reprinted in J. D. Greenwood (Ed.) (1991), *The future of folk psychology: Intentionality and cognitive science.* (pp.149-175). Cambridge: Cambridge University Press.

Humphrey, N. K. (1983). *Consciousness regained.* Oxford: Oxford University Press.

Humphrey, N. K. (1986). *The inner eye.* London: Faber and Faber.

Jackendoff, R. S. (1987). *Consciousness and the computational mind.* Cambridge, MA: MIT Press/ Bradford Books.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32,* 127-136.

Jackson, F. (1986). What Mary didn't know. *Journal of Philosophy, 83,* 291-295.

Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind.* London: Allen Lane.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (1988a). A computational analysis of consciousness. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 357- 368). Oxford: Clarendon Press. Originally published in *Cognition and Brain Theory, 6,* 499-508.

Johnson-Laird, P. N. (1988b). *The computer and the mind: an introduction to cognitive science.* London: Fontana Press.

Karmiloff-Smith, A. (1986). From metaprocess to conscious access: Evidence from children's metalinguistic and repair data. *Cognition, 23,* 95-147.

Karmiloff-Smith, A. (1991). What every cognitive psychologist should know about the mind of a child. In W. Kessen, A. Ortony, and F. Craik (Eds.), *Memories, thoughts, and emotions: Essays in honor of George Mandler.* (pp.277-288). Hillsdale, NJ: Lawrence Erlbaum.

Keat, R., and Urry, J. (1975). Realist philosophy of science. In their *Social theory as science.* London: Routledge and Kegan Paul.

Kemp, S., and Strongman, K. T. (unpublished manuscript). Consciousness — A folk theoretical view.

Kenny, A. (1984). *The legacy of Wittgenstein.* Oxford: Basil Blackwell.

Kinsbourne, M. (1988). Integrated field theory of consciousness. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 239-256). Oxford: Clarendon Press.

Kitcher, P. (1979). Phenomenal qualities. *American Philosophical Quarterly, 16,* 123-129.

Kitcher, P. (1985). Narrow taxonomy and wide functionalism. *Philosophy of Science, 52,* 76-97.

Levin, J. (1986). Could love be a heatwave?: Physicalism and the subjective character of experience. *Philosophical Studies, 49,* 245-261.

Lycan, W. G. (1979). A new Lilliputian argument against machine functionalism. *Philosophical Studies, 35,* 279-287.

Lycan, W. G. (1981a). Form, function, and feel. *Journal of Philosophy, 78,* 24-50.

Lycan, W. G. (1981b). Toward a homuncular theory of believing. *Cognition and Brain Theory, 4,* 139-159.

Lycan, W. G. (1987). *Consciousness.* Cambridge, MA: MIT Press.

Lycan, W. G. (1990). The continuity of levels of nature. In W. G. Lycan (Ed.), *Mind and cognition: a reader.* (pp. 77-96). Oxford and Cambridge, MA: Basil Blackwell.

Lyons, W. (1990). Intentionality and modern philosophical psychology, I. The modern reduction of intentionality. *Philosophical Psychology, 3,* 247-269.

Lyons, W. (1991). Intentionality and modern philosophical psychology — II. The return to representation. *Philosophical Psychology, 4,* 83-102.

Mandler, G. (1975). Consciousness: respectable, useful and probably necessary. In R. Solso (Ed.), *Information processing and cognition: the Loyola symposium.* Hillsdale, NJ: Erlbaum.

Manicas, P. T., and Secord, P. F. (1983). Implications for psychology of the new philosophy of science. *American Psychologist, (April),* 399-413.

Marcel, A. J. (1988). Phenomenal experience and functionalism. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp.121-158). Oxford: Clarendon Press.

Marr, D. (1982).*Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

McClelland, J., and Rumelhart, D. (Eds.) (1986) and the PDP research group. *Parallel distributed processing: Explorations in the microstructure of cognition. Vols. 1 & 2.* Cambridge, MA: MIT Press.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387-395.

McCulloch, G. (1988). What it is like. *Philosophical Quarterly, 38,* 1-19.

McGinn, C. (1989). Can we solve the mind-body problem? *Mind, 98,* 349-366.

McGinn, C. (1991). *The problem of consciousness.* Oxford: Basil Blackwell.

Mead, G. H. (1964). The social self. In A. J. Reck (Ed.), *Selected writings of George Herbert Mead.* (pp. 142-149). Indianapolis, IN: Bobbs-Merrill.

Millikan, R. G. (1989). In defence of proper functions. *Philosophy of Science, 56,* 288-302.

Minsky, M. (1985). *The society of mind.* New York: Simon & Schuster.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83,* 435-450. Reprinted in T. Nagel (1979), *Mortal questions.* (pp. 165-180). Cambridge: Cambridge University Press.

Nagel, T. (1979). *Mortal questions.* Cambridge: Cambridge University Press.

Nagel, T. (1986). *The view from nowhere.* Oxford: Oxford University Press.

Natsoulas, T. (1978a). Consciousness. *American Psychologist, 33,* 906-914.

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual information processing.* (pp. 463-526) New York: Academic Press.

Noble, D. (1990). Explanation and intention. In K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale, and K. V. Wilkes, (Eds.), *Modelling the mind.* (pp.97-112). Oxford: Clarendon Press.

Norman, D. A., and Shallice, T. (1980). Attention to action: willed and automatic control of behaviour. *Center for Human Information Processing Technical Report* No. 99. Reprinted with revisions in R. J. Davidson, G. E. Schwartz, and D. Shapiro (Eds.) (1986), *Consciousness and self-regulation.* New York: Plenum Press.

Oatley, K. (1988). On changing one's mind: a possible function of consciousness. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 369-389). Oxford: Clarendon Press.

Oatley, K., and Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotion. *Cognition and Emotion, 1,* 29-50.

Place, U. T. (1962). Is consciousness a brain process? In V. C. Chappell (Ed.), *The philosophy of mind*. Englewood Cliffs, NJ: Prentice Hall.

Posner, M. I., and Boies, S. J. (1971). Components of attention. *Psychological Review, 78*, 391-408.

Posner, M. I., and Klein, R. M. (1973). On the functions of consciousness. In S. Kornblum (Ed.), *Attention and performance, IV*. (pp. 21-35). New York and London: Academic Press.

Posner, M. I., and Warren, R. E. (1972). Traces, concepts, and conscious constructions. In A. W. Melton and E. Martin (Eds.), *Coding processes in human memory*. Washington: Winston & Wiley.

Putnam, H. (1960). Men and machines. In S. Hook (Ed.), *Dimensions of mind*. New York: New York University Press.

Putnam, H. (1967). Psychological predicates. In W. Captain and D. Merill (Eds.), *Art, Mind, and religion*. (pp. 37-48). Pittsburgh: University of Pittsburgh Press.

Putnam, H. (1975). Philosophy and our mental life. In H. Putnam (Ed.), *Mind, language, and reality. Philosophical papers, Vol. 2*. (pp. 291-303). Cambridge: Cambridge University Press.

Ramsey, W. M. (1989). Parallelism and functionalism. *Cognitive Science, 13*, 139-144.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329-359.

Rust, J. (1992). Editorial: Philosophical psychology in the 1990s. *Philosophical Psychology, 5*, 3-6.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417-424.

Searle, J. (1984). *Minds, brains and science: The 1984 Reith Lectures*. London: BBC.

Shallice, T. (1988). Information-processing models of consciousness: possibilities and problems. In A. J. Marcel and Bisiach, E (Eds.), *Consciousness in contemporary science.* (pp. 305-333). Oxford: Clarendon Press.

Shallice, T. (1991). The revival of consciousness in cognitive science. In W. Kessen, A. Ortony, and F. Craik, (Eds.), *Memories, thoughts, and emotions: Essays in honor of George Mandler.* (pp.213-226). Hillsdale, NJ: Lawrence Erlbaum.

Shoemaker, S. (1975). Functionalism and qualia. *Philosophical Studies, 27,* 291-315.

Shoemaker, S. (1982). The inverted spectrum. *Journal of Philosophy, 79,* 357-381.

Shoemaker, S. (1991). Qualia and consciousness. *Mind, 100,* 507-524.

Sloman, A. (1991). Developing concepts of consciousness. Open peer commentary to Velmans, M. (1991), Is human information processing conscious? *Behavioral and Brain Sciences, 14,* 694-695.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1-74.

Sober, E. (1990). Putting the function back into functionalism. In W. G. Lycan (Ed.), *Mind and cognition: a reader.* (pp. 97-106). Oxford and Cambridge, MA: Basil Blackwell. Excerpted from E. Sober (1985), Panglossian functionalism and the philosophy of mind. *Synthese, 64,* 165-193.

Sterelny, K. (1989). Computational functional psychology: Problems and prospects. In P. Slezak and W. R. Albury (Eds.), *Computers, brains and minds: Essays in cognitive science.* (pp.71-93). Dordrecht: Kluwer Academic.

Sterelny, K. (1990). *The representational theory of mind: an introduction.* Oxford: Basil Blackwell.

Stich, S. C. (1978). Beliefs and subdoxastic states. *Philosophy of science, 45,* 499-518.

Stich, S. C. (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, MA: MIT Press.

Strawson, G. (1992). The self as software: Conscious experience without the mystery. Review of Dennett's 'Consciousness explained'. *Times Literary Supplement, Aug. 21,* 5-6.

Thagard, P. R. (1986). Parallel computation and the mind-body problem. *Cognitive Science, 10,* 301-318.

Thagard, P. R. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12,* 435-502.

Thornton, M. (1989). *Folk Psychology: An introduction.* Toronto: University of Toronto Press.

Triesman, A. M. (1964a). Verbal cues, language and meaning in attention. *American Journal of Psychology, 77,* 206-214.

Triesman, A. M. (1964b). The effect of irrelevant material on the efficiency of selective listening. *American Journal of Psychology, 77,* 533-546.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59,* 434-460.

Van Gulick, R. (1980). Functionalism, information and content. *Nature and System, 2,* 139-162. Reprinted in W. G. Lycan (Ed.) (1990), *Mind and cognition: a reader.* (pp.107-129). Oxford and Cambridge, MA: Basil Blackwell.

Van Gulick, R. (1985). Physicalism and the subjectivity of the mental. *Philosophical Topics, 16.*

Van Gulick, R. (1990). What difference does consciousness make? *Philosophical Topics, 17,* 211-230.

Van Gulick, R. (1991). Consciousness may still have a processing role to play. Open peer commentary to Velman's, Is human information processing conscious? *Behavioral and Brain Sciences, 14,* 699-700.

Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences, 14,* 651-669.

Velmans, M. (1992). Is consciousness integrated? Open peer commentary to Dennett and Kinsbourne's, Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences, 15,* 229-230.

Weston, T. (1992). Approximate truth and scientific realism. *Philosophy of Science, 59,* 53-74.

Wilkes, K. V. (1978). *Physicalism.* London: Routledge and Kegan Paul.

Wilkes, K. V. (1981). Functionalism, psychology, and the philosophy of mind. In J. I. Biro and R. W. Shahan, (Eds.), *Mind, brain, and function.* Oklahoma: University of Oklahoma Press. Originally published in *Philosophical Topics, 12,* 147-167.

Wilkes, K. V. (1984). Is consciousness important? *British Journal for the Philosophy of Science, 35,* 223-243.

Wilkes, K. V. (1988a). —, yishi, duh, um, and consciousness. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in contemporary science.* (pp. 16-41). Oxford: Clarendon Press.

Wilkes, K. V. (1988b). *Real people: Personal identity without thought experiments.* Oxford: Clarendon Press.

Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind-body problem. In G. G. Globus, G. Maxwell, and I. Savodnik (Eds.), *Consciousness and the brain. A scientific and philosophical inquiry.* New York and London: Plenum Press.